# Creating a nationally representative individual and household sample for Great Britain, 1851 to 1901: the Victorian Panel Study (VPS)

Schürer, Kevin

Focus:

»Creating a Nationally Representative
Individual and Household Sample«

# Creating a Nationally Representative Individual and Household Sample for Great Britain, 1851 to 1901 – The Victorian Panel Study (VPS)

*Kevin Schürer* *

**Abstract**: This publication is a direct result of an earlier scoping study undertaken for the ESRC's Research Resources Board which investigated the potential for creating a new longitudinal database of individuals and households for the period 1851 to 1901 – the *Victorian Panel Study* (VPS). The basic concept of the VPS is to create a unique longitudinal database of individuals and households for Great Britain spanning the period 1851-1901. The proposed VPS project raises a number of methodological and logistical challenges, and it is these which are the focus of this publication.

The basic idea of the VPS is simple in concept. It would take as its base the individuals and households recorded in the existing ESRC-funded computerised national two per cent sample of the 1851 British census, created by Professor Michael Anderson, and trace these through subsequent registration and census information for the fifty-year period to 1901. The result would be a linked database with each census year between 1851 and 1901 in essence acting as a sur-

* Address all communications to: Kevin Schürer, Department of History**,** University of Essex, Wivenhoe Park, Colchester, CO4 3SQ, UK; e-mail: schurer@essex.ac.uk. For an extended version of this report cf. vol. 20 of *HSR-Transition* (www.hsr-trans.de).

The project which gave rise to this publication was funded by the UK Economic and Social Research Council and due thanks and acknowledgement are recorded. The work described here has been informed by various contributions by the project team consisting (in addition to the author) of Yijun Xue, Massimilliano Geraldi, Christine Jones and the late Alasdair Crockett. I am particularly indebted to the last of these who drafted the section on sample and weighting design. The work was also informed by the attendees to a 'User Workshop' held at the British Academy, at which initial results of the project were presented and discussed. In addition to this workshop, discussions were held with a number of experts in nineteenth-century census analysis and the construction of longitudinal surveys. These helped to both formulate ideas and correct misunderstandings. In particular I would like to thank: Michael Anderson, Ros Davies, Peter Doorn, Peter Lynn, Kees Mandermakers, Colin Pooley, Steve Ruggles, Peter Tilley, Matthew Woollard and Tony Wrigley.

rogate ‘wave’, associated with information from registration events that occurred between census years.

Although the idea of a VPS can be expressed in this short and simple fashion, designing and planning it, together with identifying and justifying the resources necessary to create it, is a complex set of tasks, and it is these which this publication seeks to address. The primary aims and objectives of the project described in this publication were essentially as follows:

- to estimate the potential user demand for a VPS and examine the uses to which it may be put;
- to test the suitability of the existing 1851 census sample as an appropriate starting point for a VPS;
- to test differing sampling and methodological issues;
- to investigate record-linkage strategies;
- to investigate the relationship between the VPS and other longitudinal data projects (both contemporary and historical);
- and to recommend a framework and strategy for creating a full VPS.
- The structure and contents of this publication follow this basic project plan.

# 1. Introduction

This publication is a direct result of an earlier scoping study undertaken for the UK Economic and Social Research Council’s (ESRC) Research Resources Board which investigated the potential for creating a new longitudinal database of individuals and households for the period 1851 to 1901 – the *Victorian Panel Study* (VPS). The basic concept of the VPS is to create a unique longitudinal database of individuals and households for Great Britain spanning the period 1851-1901. This would essentially use census and civil registration material (births, marriages and deaths) already existing in computerised form, starting with the national two per cent sample of the 1851 British census, created by Professor Michael Anderson,[1] and trace those enumerated here through subsequent registration and census information up until the 1901 census. The result would be a linked database with each census year between 1851 and 1901 in essence acting as a surrogate ‘wave’, associated with information from registration events that occurred between census years. Although relatively simple in context, the proposed VPS project raises a number of methodological and logistical challenges, and it is these which are the focus of this publication.

[1] Anderson *et al*, *National sample from the 1851 census of Great Britain*.

In this sense what this publications seeks to provide is a project plan of how a VPS might be created.

This publication describes plans for the creation of what is termed here the Victorian Panel Study.[2] The idea of generating a VPS arose from an initiative taken by The National Archives (TNA, formerly Public Record Office) to enter into collaborative agreements with appropriate HE/FE stakeholders in order to generate new ITC resources to the mutual benefit of both parties. In the discussions between TNA and the ESRC which followed, the idea of jointly creating a VPS was first raised. An outline plan was subsequently presented by the author of this publication to the ESRC's Research Resources Board who agreed to fund a preliminary scoping study. This first report (hereafter referred to as the Scoping Report) was delivered to the ESRC in October 2003.[3] On the basis of the Scoping Report an invitation was issued to make an application for further funding in order to investigate the potential for creating a VPS more fully, concentrating especially on theoretical and methodological issues. This second report was delivered to the ESRC in May 2006, and the current publication is based substantially on the work of that report.[4] Those interesting in the background to this work should refer to these two reports.

The basic idea of the VPS as proposed in the Scoping Report is simple in concept. It would take as its base the individuals and households recorded in the existing ESRC-funded computerised national two per cent sample of the 1851 British census, created by Professor Michael Anderson,[5] and trace these through subsequent registration and census information for the fifty-year period to 1901. The result would be a linked database with each census year between 1851 and 1901 in essence acting as a surrogate 'wave', associated with information from registration events that occurred between census years.

Although the idea of a VPS can be expressed in this short and simple fashion, designing and planning it, together with identifying and justifying the resources necessary to create it, is a complex set of tasks, and it is these which this publication seeks to address. The primary aims and objectives of the project described in this publication were essentially as follows:

- to estimate the potential user demand for a VPS and examine the uses to which it may be put;
- to test the suitability of the existing 1851 census sample as an appropriate starting point for a VPS;
- to test differing sampling and methodological issues;
- to investigate record-linkage strategies;

---

[2] The work on which this publication is based was funded by a grant from the UK ESRC, award reference RES-500-25-5001.

[3] Schürer, 'The Victorian Panel Survey: a scoping study for the ESRC'. Available at: http://www.data-archive.ac.uk/randd/vpsreport1.pdf.

[4] Crockett, Jones and Schürer, 'The Victorian Panel Survey: a pilot project'. Available at: http://www.data-archive.ac.uk/randd/vpsreportforrrb.pdf.

[5] Anderson *et al*, *National sample from the 1851 census of Great Britain.*

- to investigate the relationship between the VPS and other longitudinal data projects (both contemporary and historical);
- and to recommend a framework and strategy for creating a full VPS.

The structure and contents of this publication follow this basic project plan.

# 2. Potential Demand for a VPS

## 2.1. Introduction

The UK ESRC has long recognised the value of quality longitudinal data for social scientific research.[6] Indeed, one of the seven strategic objectives currently identified by the ESRC is to 'provide the data and methods needed to meet future social science challenges', which, in turn, is backed by a commitment to provide 'long term support for the UK's world-class longitudinal studies'.[7] In order to develop and guide this commitment the ESRC have also recently formulated a National Strategy for Data Resources for Research in the Social Science, which is informed by a new UK Data Forum. The resulting strategic document again highlights the need of high quality data for world-class research, noting that longitudinal data are 'particularly important, in that they allow monitoring of processes of change and facilitate research which investigates causal interpretations of these processes'.[8] However, one omission of the National Strategy is its failure to take full account of the needs of research based in social and economic history, and the potential of historical resources to provide a long-term comparative context for other areas of social science research. The creation of a VPS would aim to fill this gap.

Regardless of the technical and logistical possibilities for creating a VPS, since the main purpose in creating a VPS would be to promote and produce high quality research, it is of critical importance to assess potential user needs for a VPS in order to justify any expense that the ESRC or other bodies might make in order to create it. Thus, the project undertook a user consultation exercise, including the wider social scientific community in addition to economic and social historians, in order to try and ascertain potential demand and future research requirements from a VPS. This user consultation exercise principally took one of three forms:

- a paper and electronic questionnaire survey;
- face-to-face interviews with key individuals in the field;
- a user consultation workshop held in London.

---

[6] For example, an in-depth investigation of longitudinal studies was commissioned by the newly-formed Social Science Research Council (SSRC) in 1967. See Wall and Williams, *Longitudinal studies and the social sciences.*

[7] ESRC Strategic Plan 2005-2010, p.4 and p.17.

[8] *National Strategy for Data Resources for Research in the Social Sciences.* See http://www.esrc.ac.uk/ESRCInfoCentre/Images/National_Data_Strategy_tcm6-18160.pdf.

However, before discussing these it is instructive to examine the use made of contemporary longitudinal studies, as well as to how census and longitudinal data have been used historically.

## 2.2. The evidence of modern longitudinal studies

There are several existing longitudinal data bases covering the period from 1946 to the present.[9] These primarily consist of the British Household Panel Study (BHPS), the Birth British Cohort Study (BCS70), the Millennium Cohort Study (MCS), the National Child Development Survey (NCDS) and the ONS Longitudinal Study (LS). All of these are heavily used by academic researchers from a number of disciplines. In the reporting year 2004-05 ESDS Longitudinal, which handles distribution for all the above-mentioned databases except the ONS LS, disseminated 2,049 longitudinal datasets.[10] The disciplinary spread of this demand for data was as follows: Economics and labour studies, 44%; Sociology, 17%; Social Policy, 8%; Health, 7%; Statistics, 6%; Geography, 4% and Psychology, 3%.

The provision of these contemporary longitudinal datasets has generated a large number of research publications and related outputs. The online publication databases related to these studies collectively list some 3,229 publications to date.[11] It is also the case that these publications span a broad area of disciplines, topics and issues. This can be illustrated by the research output arising from the ONS LS which can be classified according to the subject groups or themes given in Table 2.1.

Thus the provision of quality longitudinal data for more modern periods has generated a wealth of research outputs across a broad range of social science (and related) disciplines, thus justifying the significant investment made by the ESRC each year in supporting the generation of such longitudinal data.

## 2.3. Use of historic census data and longitudinal data

The manuscript nineteenth-century census returns, usually referred to as the census enumerators' books (CEBs), have previously been heavily used in historical research. Indeed, they could be said to form the backbone of many strands of investigation within social and economic history for the nineteenth century. An annotated bibliography of research based substantially on the

[9] A database on longitudinal data resources is maintained by the UK Longitudinal Studies Centre, called *Keeping Track*. See <http://www.iser.essex.ac.uk/ulsc/resources/keeptrack.php>.

[10] For more information on ESDS Longitudinal see <http://www.esds.ac.uk/longitudinal/introduction.asp>.

[11] See <http://www.celsius.lshtm.ac.uk/publications.html>; <http://www.iser.essex.ac.uk/pubs> and <http://www.cls.ioe.ac.uk/publications.asp?section=000100010006>.

CEBs published in 1989 lists some 423 publications.[12] These can be subdivided into the thematic categories listed in Table 2.2. In the fifteen years that have passed since this list was compiled the number of published studies has clearly increased, since the popularity of the CEBs as a key source shows no sign of diminishing. The current figure of published outcomes from the CEBs is now possibly double that reported in 1989, at around 850, but there is no reason to suggest that the overall balance in terms of subject matter is fundamentally different from that recorded in Table 2.2. It is also important to note that these figures take no account of unpublished masters and doctoral research, for which the CEBs have been extensively used within the field of nineteenth-century social and economic history.

Tab. 2.1: Subjects researched using the ONS LS

| *Category* | *n. of publications* | *%* |
|---|---|---|
| Ageing | 52 | 8,4 |
| Births | 50 | 8,1 |
| Cancer studies | 91 | 14,7 |
| Death | 204 | 32,9 |
| Environmental | 10 | 1,6 |
| Ethnicity | 78 | 12,6 |
| Fertility | 31 | 5,0 |
| Gender | 94 | 15,1 |
| General methodological | 91 | 14,7 |
| Geographic differences in health/mortality | 26 | 4,2 |
| Health | 65 | 10,5 |
| Household change | 62 | 10,0 |
| Housing | 108 | 17,4 |
| Inequalities in health and mortality | 129 | 20,8 |
| Infant mortality | 3 | 0,5 |
| Local geography | 147 | 23,7 |
| Migration | 87 | 14,0 |
| Mortality | 225 | 36,2 |
| Occupational mortality | 8 | 1,3 |
| Social and economic change | 71 | 11,4 |
| Social mobility | 219 | 35,3 |
| Still births | 2 | 0,3 |
| Widowhood | 12 | 1,9 |
| Total | 621 | 100 |

Note: The numbers do not sum to the total since some publications are classed in multiple categories.

[12] Mills and Pearce, *People and places*, 6. However the number of separate publications is a little under 400 since some publications are listed more than once due to the nature of their geographical coverage.

Tab. 2.2: Topics investigated in the CEB publications

| | | |
|---|---|---|
| Occupations | 257 | 60,8% |
| Migration | 218 | 51,5% |
| Demographic | 168 | 39,7% |
| Household | 160 | 37,8% |
| Methodology | 122 | 28,8% |
| Segregation | 100 | 23,6% |
| Social structure | 69 | 16,3% |
| Family | 66 | 15,6% |
| Total | 423 | 100% |

Note: The numbers do not sum to the total since some publications are classed in multiple categories; calculated from Mills and Pearce, *Peoples and places*.

As with research on modern day populations, there are many potential research questions that historical longitudinal datasets can address which cross-sectional data, such as recorded in the CEBs, simple cannot. Indeed, it is because of the potential offered by longitudinal data that a number of researchers have already constructed longitudinal databases based on the CEBs. Research using linked CEBs has already made important contributions and offered valuable insights into a number of areas of social and economic history. These have included work on historical demography,[13] migration,[14] occupational structures,[15] household composition,[16] as well as evaluations of the accuracy of the underlying source materials.[17] However, the existing body of nineteenth-century longitudinal data using the CEBs, and the studies arising from them all suffer from two common and fundamental methodological problems, which the VPS would seek to overcome. First, out of practical necessity they are all essentially 'place orientated', based on local populations. Previously, the basic method of constructing a longitudinal database from the CEBs has been to take the records of individuals and households for one particular place (usually a parish) and to link them incrementally to the CEBs for subsequent censuses for the same place. As a result a large number of individuals are lost from observation (potentially up to forty per cent, and often more in large urban areas) as they move away from the place being studied between censuses. Thus the

[13] See, for example, Hinde, 'Population of a Wiltshire village'; Garrett, 'Trials of labour'; Reay, 'Before the transition'.
[14] See, for example, Dennis, 'Distance'; Dennis, 'Intercensal mobility'; Pooley, 'Residential mobility'; Pooley and Doherty, 'The longitudinal study of migration'; Schürer, 'The role of the family'; White, 'Family migration'; Wojciechowska, 'Brenchley'.
[15] See, for example, Crompton, 'An exploration'; Hallas, the social and economic impact'.
[16] See, for example, Hancock, 'In service'; Nenadic, 'Studying the middle class'; Reay, 'Kinship and the neighbourhood'.
[17] See, for example, Perkyns, 'Age checkability'; Perkyns, 'Birthplace accuracy'.

resulting database and research results derived from it are essentially those for a stable or non-migratory subset of the population.[18]

Second, since access to the civil registration material has not been available in a useable and effective form, previous studies constructing longitudinal databases have tended to link census to census. This not only reduces the number of links possible,[19] but also means that intervening events are outside of observation. The inability to link the civil registration data to the CEBs also reduces the range of information that can be associated with an individual's 'life history' record.

## 2.4. Potential Uses for a VPS

The user survey, face-to-face meetings, and the consultation workshop, together with the review of how existing longitudinal databases and historical census data have been used, combined to provide an insight into the potential uses to which a VPS might be put. These are multi-disciplinary and multi-faceted. It is also clear that the ideas put forward embrace research questions that simply cannot be addressed using existing cross-sectional or local sources.

It is clearly not possible to predict all the areas of research that the VPS, if created, might be used to address, but the following thematic sub-sections provide an indication of potential areas of investigation.

### *2.4.1. Demography*

- The analysis of demographic events (fertility, nuptiality and mortality) within the dynamic household and familial contexts in which they occur.
- The construction of cohort demographic rates in order to re-evaluate existing knowledge of the demographic transition, most demographic work for the nineteenth century previously being based on period rates.
- The construction of cohort life tables by occupation groups and social class.
- Contrasting the demographic experience of migratory and non-migratory populations.
- The analysis of demographic experience by religious denomination.
- The analysis of cause of death within the context of dynamic household and family structures.

[18] A similar problem, of course, exists in the case of family reconstitutions conducted for earlier periods using parish registers in order to calculate historical demographic rates. See Wrigley *et al*, *English population history*.

[19] In particular, single women may be lost from observation following marriage due to their change of surname, and children who are both born and die between censuses are lost. Researchers have used local parish registers to overcome these problems, but due to the incompleteness of ecclesiastical registration for most places in the second half of the nineteenth century, this offers only a partial solution at best.

- The investigation of generational effects on nuptiality, fertility and mortality (the extent to which a parent's demographic experience, influences that of their children).
- The analysis of demographic experience by literacy/illiteracy.

### *2.4.2. Migration*

- The reconstruction of macro-migration experiences over the life-course in the context of the timing of demographic events.
- The analysis of regional migration flows and changing regional economic structures, in a socio-economic context.
- Changes in house occupancy (in a particular street, for example, if area-based sub-sampling is used).

### *2.4.3. Emigration*

- The construction of emigration rates and estimates.
- The analysis of the demographic, social and economic profile of emigrants.

### *2.4.4. Immigration/Overseas born population*

- Creation of specialised sub-samples of the overseas born population (to resolve refreshment issues) will enable detailed analyses of the characteristics of the immigrant population in this period.

### *2.4.5. Household and family*

- The investigation of the processes of household evolution and dissolution within regional, social and economic contexts. The examination of non-family members (servants, lodgers, boarders) and their movement into and out of 'normal' family groups.
- The analysis of household structures by religious denomination.
- The examination of individuals moving to and from familial patterns of residence to institutional living arrangements (e.g. workhouses).

### *2.4.6. Employment*

- The analysis of occupational mobility – father/son occupations at different dates.
- The analysis of social mobility – father/son occupations at different dates.
- Examination of career profiles – e.g. retirement
- The examination of changing household-based work histories.
- The validity and consistency of occupation recording.

### *2.4.7. Literacy and language*

- The examination of literacy in the context of generational, regional, social and economic indicators.
- The investigation of Gaelic and Welsh speaking populations in a longitudinal context.

### *2.4.8. Social processes*

- The analysis of the experience of change over time for particular sub-groups of the population. For instance: particular occupational groups, ethnic (birthplace) groups, those at a particular stage of the life cycle, those in particular household positions. Thus it would be possible to focus (for example) on married women age 20-29 and examine their occupations and family circumstances over a 50-year period.
- Socio-economic change in relation to poverty.

### *2.4.9. General*

- The examination of regional change with respect to (for instance) occupations, household structure, birthplace/ethnicity.
- The development of vignettes, or life histories, of particular groups within the population as illustrations of broader trends.
- The investigation at a national and regional scale of a whole range of questions that are usually only investigated at a local scale. For instance, examination of changes in detailed household structure with regard to the position of servants and lodgers.
- Methodological research into record-linkage, census enumeration.

The VPS would provide an important research resource across a range of disciplines, not only history, and could be used to address a broad spectrum of research questions that cannot be easily or satisfactorily answered by other existing sources. The creation of a VPS would also add to a growing number of historical longitudinal studies, broadening and facilitating cross-national comparative on the topics listed previously in this section.

# 3. The 1851 Census national sample

## 3.1. Background

The original concept of the VPS was that it would take as its base and starting point the existing two per cent sample of the 1851 census for Great Britain, created by Prof Michael Anderson and his associates at Edinburgh University

during the 1970s.[20] This assumption was made essentially on the grounds of efficiency and effectiveness – it being quicker and cheaper to start the proposed panel study from an existing statistically representative sample, rather than create an entirely new sample drawn from the 1851 census. However, before confirming this as a recommendation to the project, it was necessary to determine that using the existing 1851 sample is the optimal solution. Alternatives needed to be explored, and equally, the validity and properties of the Anderson sample as a basis of a representative longitudinal database needed to be tested.

Following discussions with the database creator, Michael Anderson, it was decided to undertake in-depth checks on the 1851 sample, focusing on a number of issues:

- the quality of the transcription;
- the enumeration of vessels;
- the enumeration of institutions.

### *3.1.1. Transcription quality*

In order to test transcription quality a simple but time-consuming course of action was taken. A sample of paper photocopies of the original source material was selected for 54 parishes (covering 37,768 individuals, about 9 per cent of the entire sample) and compared, line-by-line, variable-by-variable with the machine-readable version that has been created by Anderson and his team. The result of this exercise, suggest that the standard of the transcription was generally very good, particularly considering the poor quality of some of the original CEBs – generally relationships, marital condition, gender and ages were transcribed very accurately. Overall, the exercise revealed the following discrepancies:

- 76 individuals (0.2 per cent of the total) were missing from the transcription;
- 5 individuals (0.01 per cent of the total) were duplicated in the transcription;
- 46 entries for individuals (0.12 per cent of the total) were 'spurious' (that is, were for individuals who did not actually exist);
- 4 entries for individuals (0.01 per cent of the total) were transcribed incorrectly.

However, in undertaking a series of more general checks, checking parish and 'cluster' totals[21] across the machine-readable version of the sample a number of more substantial errors were noted. This, rather surprisingly, brought to light that three enumeration districts and an institution were missed out of the

[20] Anderson, *et al, National sample from the 1851 census of Great Britain.* For details of Prof. Anderson's 1851 census sample see *ScopingReport*, 19-21.

[21] Cluster is a term used in the documentation for the 1851 census sample. It broadly relates to a census enumeration district, with the exception of institutions, the samples of which are also termed clusters.

machine-readable version of the sample, while both an enumeration district and an institution were duplicated. As a result the machine-readable sample has subsequently been corrected.

Since surnames and forenames are key fields to be used in any record linkage exercise, it was also decided to undertake a detailed investigation of the quality of their transcription within the 1851 sample database. This revealed that, perhaps worryingly, surnames were totally absent for 1.6 per cent of the sample and partially incomplete for a further 3.8 per cent. Since this could have a direct impact on the ability of perform successful record linkage on these individuals, further analysis was undertaken to test if missing or incomplete name information was random across the database, or might be a risk in introducing bias into the linked database. This exercise initially identified those working in the textile industries as have significantly higher levels of incomplete name data, with textile and fabric workers being three times as likely to have absent name information than the 1851 sample as a whole. Slight increased levels of incomplete name data were also found for those working in Domestic Service and Agriculture. In terms of relationship to household head, patients and inmates in residential institutions records higher incidences of missing name information, but this is to be expected given the way in which those resident in mid-nineteenth institutions were enumerated.[22] Interestingly, those enumerated in Wales and Scotland both had a higher level of name completeness that those enumerated in England (Scotland 98.3% complete; Wales 97.8%; England 92.4%), while comparing the British born with those born overseas, revealed, perhaps surprisingly, but encouragingly, that there is no apparent increase in incompleteness of name information for immigrants – both groups a rate of 93.5 complete.

However, further investigation subsequently showed the link between those employed in textiles and name incompleteness to be somewhat spurious. When examined in relation to the region of enumeration, this revealed that there was nothing peculiar about textile workers *per se* regarding name incompleteness, but rather it was being enumerated in Lancashire or Cheshire that was the significant factor. It is only because textile workers were so concentrated in these two counties (together they account for 38% of all textile/fabric workers) that this occupational group has higher rates of incomplete name information, Indeed, outside of Lancashire and Cheshire textile workers display slightly higher complete rates of name information than the sample as a whole. Instead, the exercise revealed that a relative small number of CEBs within the Anderson sample had been transcribed with a lot of blank entries across a number of fields: not just forename and surname, but also parish, county and country of birth. Subsequent research revealed that this resulted from the fact that when the photocopies from which the transcription was made were produced in 1976, the archival microfilms from which they were reproduced were of poor quality.

[22] Higgs, *Clearer sense*.

It was also discovered that the original archival census microfilms available at the time of the Anderson project were replaced with improved versions in 2003. These new films were inspected by the VPS team and it was found that, although still difficult, the microfilms are no longer impossible to read. Consequently, a new transcription was made to replace previously unreadable copies, thus significantly reducing the number of records with missing or incomplete name information.

### *3.1.2. Vessels*

The 1851 sample contains no information on ships or vessels enumerated in coastal waters since it had been assumed by Anderson and his team that the original 'shipping' returns had been lost. However, subsequent literature on the subject is ambiguous about the survival, or otherwise of the 1851 census returns for vessels.[23] Searches at TNA confirm that there has been no rediscovery of schedules for vessels, and research into the enumeration process in 1851 suggests that given the way it appears that most of the floating population were enumerated it would not be possible to re-sample this element of the population even if the schedules or returns were re-discovered.

The main question, then, for the VPS is to what extend would this missing element of the population impact on the representativeness of the overall study. In her study on the floating population, Burton suggests that it was 'highly untypical of the population at large in terms of gender, age, birthplace and occupation'.[24] Given the missing information for 1851 this impression must be based on later years, it is possible to examine the available census data for 1881 to estimate what the missing 1851 floating population might have looked like. The age/sex distribution of the population enumerated on vessels in 1881 is given in Figure 3.1.

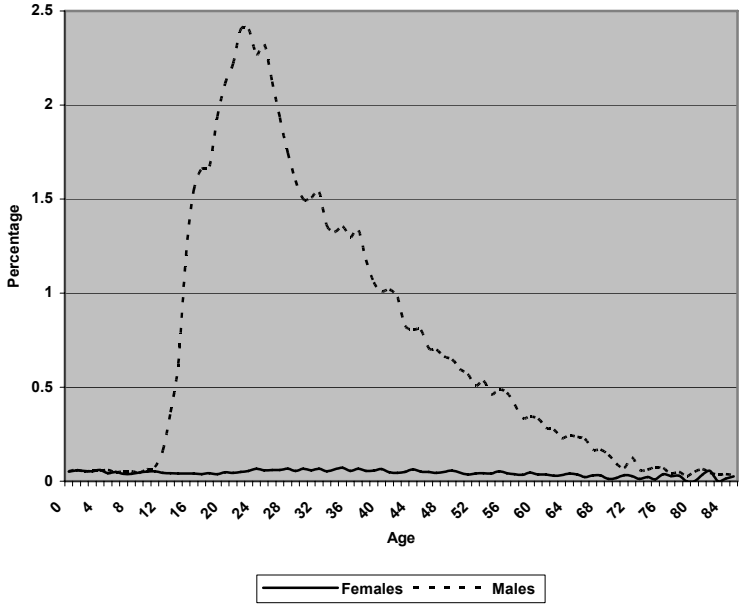
Given the near total absence of nominal level census data for those onboard vessels in 1851 there is not much that can be done to rectify this. It is impossible to go back to the original CEBs and resample the floating population as the data no longer exist and in some cases never existed in anything other than aggregated tabular form. It seems that there are two basic possibilities. First, to treat this a bit like a refreshment problem (see section 4.3 below), and add in a sample of those on vessels from 1861 onwards, from which point the data are available at the nominal level. Second, simply leave the floating population out of the panel study all together. This may seem a bit harsh, but good precedents exist: there is no comparable population, for example, in either the present day Longitudinal Study or the BHPS. Equally is must be remembered that the floating population only accounted for 0.31 per cent of the Great British total, of which 0.06 per cent were enumerated on inland barges.

---

[23] See Burton, 'A floating population' and Higgs, *Clearer sense*, 38-45

[24] Burton, 'A floating population', 49.

Fig. 3.1: Age/sex distribution of the 'floating' population: 1881




Note: this includes both those on board vessels in coastal ports and inland river craft.

### *3.1.3. Institutions*

For the 1851 sample the selection of institutions (for example, schools, workhouses, prisons, barracks) had been treated differently from private households, with only the very smallest of institutions being included in their entirety. It was therefore wondered if the way in which institutions had been selected may introduce a bias. The sampling process used for selecting institutional records for 1851 was as follows: first, individual institutions were identified from the footnotes to the population tables in the published report (in all there were 2,017 institutions enumerated in 1851).[25] Second, from these a running total was made of the numbers of people enumerated in institutions. Third, the institutions were arranged as if enumerated consecutively and, starting from a random number, 20 successive people from every 1,000 were drawn to form the sample, with the exception that families should not be split. In at least one large institution 2 blocks had been drawn. This method of sampling from the institutional population would appear to be fine, as long as there was a not a systematic selection bias in the way in which individuals within institutions were

[25] BPP 1852-3 LXXXV, xliv, Table XX.

enumerated, such as recording the inmates in a workhouse by sex and then age order, or soldiers in a barracks by order of rank (which could also be age-related). Given that some researchers had previously pointed to those within institutions sometimes being enumerated in seemingly peculiar order (age, alphabetically, etc) this would need to be checked.

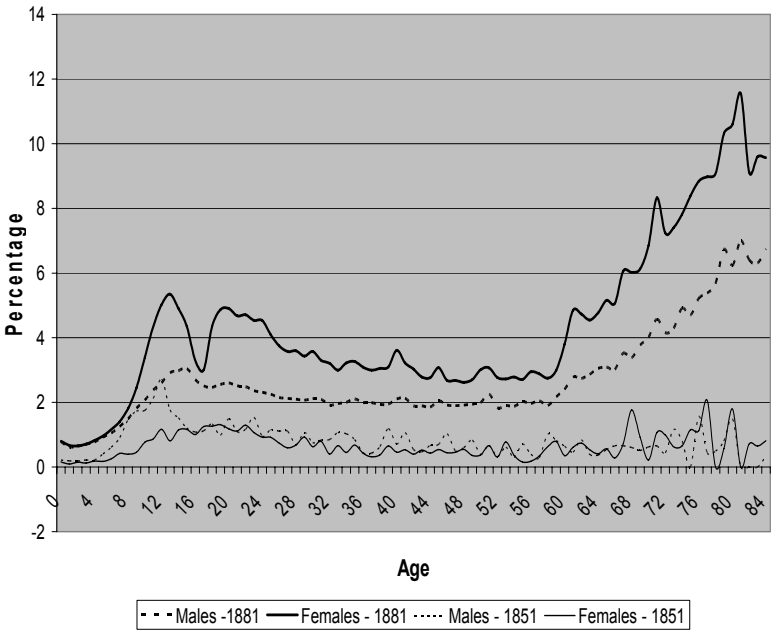
It is possible to test this methodology using the 1881 census data, drawing a sample from the complete machine-readable version in the same way and comparing the two. This exercise confirmed the basic soundness of the method, since the 1881 sample drawn by this method could be demonstrated to be representative of the complete institutional population. However, this still leaves a number of oddities with the institutional population recorded in the 1851 sample database unexplained. Comparing the age/sex distributions of the institutional populations in 1851 and 1881, having first adjusted the 1881 to take out those enumerated on vessels, as illustrated in Figure 3.2, shows a marked disparity between the two. Not only is the overall proportion of males and females within the population as a whole who were enumerated as resident in institutions rather greater in 1881 than it was in 1851, but the age structure also appears to change over time. In 1881 the proportion of females resident in institutions rises sharply from the age of 60, increasing from around 3 to 10 per cent of the total population by the age of 85. A similar increase occurs for males, but at a slightly lower level, from around 2 to 7 per cent. However, no such increase occurs in the 1851 Anderson sample population, with the rate of both males and females resident in institutions between ages 60 and 85 fluctuating between 1 and 2 per cent.

It is certainly the case that structural differences in the institutional population are likely to have occurred between 1851 and 1881. For example, there may have been changes in the number size and structure of those living in military establishments. Likewise, following the so-called Groschen Minute of 1870 the central Poor Law Board advocated a general tightening-up of out-door relief, in favour of in-door relief. Indeed, Thompson has demonstrated that nationally the proportion of elderly within the workhouse population rose between 1851 and 1891.[26] However, even these changes would probably only go someway in explaining the volume of difference shown by Figure 3.2. It also totally fails to explain why there are fewer individuals within the Anderson sample recorded as living in institutions than one would expect. The 1851 Report gives a figure of 293,201 as living in institutions (excluding vessels)

[26] Thompson, 'Workhouse to nursing home', see especially Table 1, p. 49. However, it would appear that the national trend was not experienced throughout the country. For example, Jackson, 'Kent workhouse populations' shows that the 1881 situation in Kent was little changed from the mid-nineteenth-century pictures provided by Hinde and Turnball, 'The population of two Hampshire workhouses' and Goose, 'Workhouse populations in the mid-nineteenth century'. See also Jackson, 'The Medway Union Workhouse'. There is also evidence to suggest that some local Unions deliberately ignored central government policy on the nature of relief. See Hurren, 'Welfare-to-work' and Mackinon, 'English Poor Law policy'.

within Great Britain.[27] Of this, one would expect a 2 per cent sample population of 5,864, but the Anderson sample file only includes 4,763 as living in institutions, a shortfall of some 20 per cent.

Fig. 3.2: Age and sex distributions of institutional populations, 1851 and 1881




Because of these unexplained differences a detailed check was made on the institutional population in the sample using the Anderson team's original photocopies of the pages of the CEBs. This suggested that the selection of records for the sample is not always obvious. In many institutions only inmates have been drawn into the sample. These are often, but by no means always, ordered or segregated by gender, age, rank or by first letter of surname (but rarely in full alphabetical order). People of the same surname are sometimes, but not always, enumerated together. Children sometimes appear to have been enumerated with their mothers. Where officials are included in the sample their families and servants are often also present. In the case of hospitals, nursing staff are sometimes scattered through the wards. In barracks junior officers may be with their companies. The exercise also revealed that individuals for the sample

[27] BPP 1852-3 LXXXV, xliv, Table XX. Thus figure excludes the Islands in the British seas.

were not always drawn consecutively from the CEBs in blocks of twenty, as the sampling rules suggested.

As a result of these largely unexplained problems with the institutional population with the Anderson sample, it is recommended that should the Anderson sample be used then the institutional population is re-sampled, either using the method originally proposed by the Anderson team, or taking a sample of 'whole' institutions, or a combination of both.

## 3.2. Alternatives to the Anderson sample

The initial assumption that the Anderson 2 per cent 1851 sample would be used as the starting point for any subsequent VPS was based largely on the fact that no obvious rival candidate dataset existed for 1851. This, however, is no longer the case. Machine-readable versions (albeit not complete in terms of the fields covered) now exist for Scotland, England and Wales.[28] Although access to these new data collections is certainly not guaranteed, should access be possible, this raises the prospect of drawing a new sample of 1851 census data and using this as an alternative to, or in combination with the Anderson sample. This possibility is discussed further in section 4.2 below.

Other possibilities were also suggested during the course of the project. One local/family historian accused the project team of 're-inventing the wheel', claiming that so much work had already been undertaken by local historians linking data from multiple sources over time that the proposed project would be wasting its time linking data on such a large scale but rather should use what had already been created, arguing that the information required for a VPS is 'held on the databases of hundreds, if not thousands, of family historians'.[29] Likewise, genealogists have suggested that one could collect together a number of pre-existing and linked family histories or pedigrees and use these as a basis for a longitudinal dataset

Although it is certainly true that much data has been created by others, especially those working in family history, and fruitful collaborations between academic and local/family historians have been forged in the past,[30] it is believed that relying on linked data produced by others would be rather problematic in the case of the VPS.[31] Assuming that one could even collect the data

[28] See http://www.nationalarchives.gov.uk/census/ and http://www.scotlandspeople.gov.uk/.

[29] Letter from John Pollock, Family and Community Historical Research Society Newsletter 6/2 (2005), p.15.

[30] The classic example would be the work of the Cambridge Group for the History of Population and Social Structure in collecting aggregate parish register information which lead to Wrigley and Schofield's, *Population History of England*. See also Pooley and Turnbull, *Migration and Mobility*.

[31] See response to Pollock *Family and Community Historical Research Society* newsletter 6/3 (2005), p.11.

from local/family historians,[32] stitching these together in an aggregated database would not be straight-forward. It would certainly not result in a statistically representative national longitudinal database as is envisaged under the VPS, and all kinds of complex weights would need to be devised and applied in an attempt to make it representative. The problem of the need for refreshment of the underlying sample due to changes in the base population structure would be particularly taxing. The resulting database would also be methodologically flawed since different contributors would have used different criterion to decide what constitutes a 'correct' link, and also different sources would have been used in different studies thus further biasing the likelihood of linking, and consequently rates of attrition. Even if it were technically possible to overcome these methodological problems, because of the inherent complexity of creating a representative sample from pre-existing datasets using a range of differing sources and methodologies, it is suggested that ironically it may actually be less resource intensive to create a new sample afresh without all the sampling and other methodological problems attached.

# 4. Sample design, Rules of Observation and Refreshment

## 4.1. Introduction

Two of the most important tasks for the research project on which this publication is based were to explore the feasibility and sampling implications of differing observation rules, and inter-related to this issue, to examine the needs of and problems associated with differing data refreshment strategies. Impacting on both of these issues are the potential problems associated with the cluster nature of the 1851 Anderson sample. Unlike most other longitudinal studies, for pragmatic reasons at the time, the basic sampling unit for the Anderson 1851 sample was whole enumeration districts, adding an important geographical dimension to the sample. Thus, in considering the two issues of observation and refreshment, it also became important to investigate the whole issue of sample design.

## 4.2. Sample design

The most important objective of a future VPS project will probably be to create a sample that is representative of the British population – both cross-sectionally at each census year from 1851 to 1901, and in longitudinal terms over the pe-

[32] In our response to Pollock, a plea was also issued asking volunteers to to submit details of the linked data they had in their procession and would be willing to make available. The plea met with only one response.

riod 1851 to 1901.[33] The methods of achieving such a sample are discussed in detail below. Such methods should make the VPS sample more sophisticated and methodologically rigorous than any British historical dataset to date. Within the confines of what can be achieved with historical data, it is intended that the VPS sample will follow best practice of the most sophisticated contemporary longitudinal household panel surveys such as the British Household Panel Survey (BHPS) and the German Socio Economic Panel Survey (GSOEP)[34].

A key assumption underlying this project has been that the initial starting point for the VPS would be the 1851 Anderson sample (or a sub-sample thereof). However, a key question that needs to be addressed is should a fresh sample from the 1851 census be drawn, according to a new design. Cost constraints would practically rule out the possibility of making a new random sample using the original CEBs, however, it may be possible to obtain a full transcription of the census from a commercial company.[35] While the existing Anderson sample is seen as fit for VPS purposes, there are potential advantages to having a full 1851 computerised transcription available, both to allow the VPS project to define its own sample design and to trace immigrants between 1851 and 1861.

If a fresh sample is drawn, thought would need to be given as to the best methods. In contemporary surveys, equal probability selection methods (EPSEM) are seldom used, but this is usually because of cost considerations. Other methods cause design effects which should (but often are not) be corrected for in secondary analysis, whereas EPSEM samples can be analysed with no such adjustments, and results will have the correct precision.

The strategies of *clustering* and *stratification* are often used in contemporary surveys. Geographical clustering means only selecting people (or households, addresses, etc.) from a selection of geographical areas rather than from all of them. The result is sampling at quite a high rate from the minority of selected areas and at a zero rate from the majority of unselected areas. This makes the geographical unit, typically a quite small one such as postcode sector the primary sampling unit (PSU) as opposed to the person (or household, address, etc.) as it is typically in the case of an EPSEM sample. Stratification is another

[33] In the longitudinal sense, no sample can be representative of the changing British population between 1851 and 1901, due to immigration. In technical terms, the VPS sample should aim to be representative of how the British population as it constituted at 1851 changed over the 1851 to 1901 period. In a longitudinal sense, the VPS cannot reflect the increasing immigrant population over the 1851 to 1901 period, since nothing is known about immigrants prior to the first census in which they were recorded. In a cross-sectional sense the VPS can and will be representative of the growing immigrant population due to the refreshment rule outlined at section 6.3.2.

[34] See www.diw.de/english/sop. The GSOEP was started in 1984. Twenty years on, the survey consists of around 12,000 households and about 22,000 persons.

[35] For example the 1851 census returns for England and Wales have already been computerised by Ancestry.com.

very commonly used method. This involves ranking sampling units according to certain important observed characteristics (typically socio-demographic) and selecting every *n*th PSU, which increases the representativeness of the resulting sample with respect to these characteristics (compared to selecting sampling units at random). The result of stratification is that the sample size can be substantially smaller than an EPSEM sample but achieve the same statistical precision (i.e. resulting population-level estimates will be as precise), this is encapsulated in the concept of the *effective sample size*.

For these reasons, both clustering and stratification are highly cost advantageous strategies for achieving the desired effective sample size for contemporary surveys, and hence some form of stratified cluster sample is often used (i.e. both methods are employed).

It is important to realise that cost consideration would also be important for an historical survey making a sample from paper returns and computerising the selected returns post selection, as was the case with the Anderson sample. Selection and data entry will be far easier (and hence cheaper) if blocks of returns are selected (i.e. clustering). However, if the VPS project gains access to a full machine-readable version of the entire 1851 census then a practical option would be to draw a new EPSEM 1851 sample with the household as the PSU. This would also allow the possibility of producing a higher sample size, say 5 per cent rather than 2. Thus, should it be made possible through the availability of a new machine-readable version of the 1851 census, the VPS project needs to weigh up the advantages of an EPSEM sample with no design effects versus a non-EPSEM sample in which design effects can be controlled for via analysis.[36]

Whatever method is chosen, the sample should be household based, since maximal household information maximises the research utility of the data, as is recognised in contemporary longitudinal survey design (BHPS, GSOEP, etc.). Themes such as household formation over time, marriage markets and changes in patterns of domestic service over time are all better addressed with full household information. Gathering information on whole households, therefore, needs to be part of the VPS sample design, while replacement and observation rules need to ensure that maximal household information will exist for all sample members from 1851 to 1901.

It should be noted that the definition of household applied by the Victorian census authorities is different and more loosely defined than by contemporary social scientists. More problematic, it also changes over time. The issue of what constitutes a household in census returns and whether any inconsistencies are likely to affect the VPS needs to be considered.[37] Special consideration also needs to be given to the non-household (institutional) population, to insure this

[36] The possibility of taking a new EPSEM-based sample of 1851 was discussed at the consultation workshop and was widely supported by the participants, including Michael Anderson, the originator of the existing 1851 sample.

[37] Higgs *Clearer sense,* 65-71; Mills and Schürer, *Local Communities*, 281-4

is sampled in the same proportion as the household population (see section 3.1.3).

A further point to note is that the sample would be based on households as constituted in 1851. In particular, children of sample members born prior to 1851 but not resident in the household of an 1851 sample member would not be VPS sample members. Back linkage to trace such children born between 1837 (start of Civil Registration) and 1851 (but who had left that household by the 1851 census) may be desirable to provide supplement information, but should not be seen as absolutely necessary.

### *4.2.1. Sample size*

Many important research questions can only be answered if the data track substantial numbers of individuals (and their descendants) over time, as opposed to having large samples at every ten year census point, but relatively few individuals in common between the censuses. Of particular importance in the longitudinal context is the number of individuals for whom maximum longitudinal information is available, as opposed to the cross-sectional sample size at any given census year. In contemporary surveys this is known as the 'constant panel' size, this being the number of individuals responding at each survey point. In the VPS the constant panel would equate to all individuals traced for all census years from 1851 to 1901. This will inevitably be quite small in relation to the cross-sectional VPS sample size at any given year because of mortality and emigration over the 50-year period. The important data quality issue will be what percentage of individuals not in the constant panel exited, or were lost from observation due to reasons other than death or emigration. In longitudinal surveys losing individuals from observation in this way is normally termed attrition. Importantly, if attrition due to failed record linkage is common, this may bias the dataset (see section 4.4.3.1 below).

Since tracing inter-generational relationships and characteristics will be an important aspect of the VPS, a more relevant yardstick of data utility will be the success in tracing children (and grandchildren) of sample members. Linkage to civil registration birth records should provide a comprehensive initial linkage of children to parents, the less certain issue is what percentage of children will be traceable in the census, particularly censuses straddling the event of leaving home.

Thus the size of the 1851 sample is an issue of fundamental importance, the cumulative magnitude of *attrition* determine what sample size at 1851 is required to produce both large enough cross-sectional samples through to 1901 and, more importantly, a large enough longitudinal sample size – the degree to which all individuals can be traced through to death/emigration, and to which full inter-generational linkage can be made.

Both the cross-sectional sample size of latter census years and the longitudinal sample sizes (however defined) rest upon the rate of attrition, due to failed record linkage, between 1851 and 1901. This issue is considered below.

### *4.2.2. Attrition and sample size*

This rests upon the number of persons in the 1851 sample that cease to be observed (for reasons other than death or emigration) between 1861 and 1901.[38] Unlike with contemporary social surveys, the rate of attrition is likely to be quite linear over time. Contemporary surveys typically experience high attrition both when the survey commences (in terms of the initial response rate being less than 100 per cent), and between the first and second wave or sweep of data. After this attrition rates tend to decline: people who respond twice are very likely to stay in observation thereafter. In the case of the VPS, attrition rates are likely to slow down only slightly. They are likely to decline a little as those lost from observation between 1851 and 1861 are likely to have non-random characteristics (such as having common combinations of forename and surname, being more likely to migrate, etc.).

Owing to considerable uncertainty as to the attrition rate between any two census years and the cumulative attrition rate between 1851 and 1901, this project cannot make any specific recommendations regarding what sample size is needed in 1851 to achieve a 1901 cross-sectional and longitudinal samples of sufficient size. However, this is not problematic because a major advantage of this form of retrospective survey design is the ability to increase the sample size iteratively until the desired size is achieved. Despite the underlying importance of the Anderson 2 per cent sample, the 1851 sample size does not need to be irrevocably predetermined. If the initial 1851 sample proves too small, the VPS project can easily augment the 1851 sample (following the same sampling rules) in order to obtain a sample of adequate size. Such flexibility is in contrast to contemporary longitudinal surveys when the sample size that can be observed for the duration of the survey is inevitably fixed by the size of the first survey's sample.

More generally, and thinking of the VPS as a long-term historical resource that may have several waves of funding, and potentially be added to by different research groups, the main VPS sample could be progressively augmented, given sufficient thought was given to following the same rules, applying the same record linkage algorithms and to subsequent recalculations of weighting/grossing variables.

[38] Emigration cannot be traced at the individual level, but aggregate estimates between census years can be made, and used to discount a certain proportion of those who cease to be observed between census years and whose death cannot be traced, and thereby provide an estimate of numbers ceasing to be observed because of record linkage failure.

### *4.2.3. Sampling to minimise multiple links*

In the discussion of record linkage which follows, it is proposed that it may be appropriate to devise an alternative EPSEM-based sample which takes into account the forename-surname combination of heads of household. The logic of this would be to omit households from the sample with a high probability of multiple linkage problems. This novel proposal is discussed in detail in section 5.3.3. below. It is believed that this could be achieved without biasing the underlying sample, and it is recommended that this alternative approach to sampling be investigated as part of a re-sampling exercise should the issue of unresolved multiple links in the record linkage process prove to be overly problematic.

Linked to this method of using frequency of forename-surname pairs as a sampling criterion, it is also worth noting that some historical longitudinal studies have used surnames as a means to select the sample records. For example the *Base TRA Patrimoine* research project situated in Paris, is a longitudinal sample of Frenchmen dying between 1800 and 1940 whose surname starts with either the letters T, R, or A.[39] Likewise, a proposed research project to construct a longitudinal database for Scotland, 1855-2001 also intends to use surnames as a basis for sample selection.[40] This approach, however, is rejected since it is potentially open to sample bias (surnames are not evenly distributed geographically) and would make refreshment almost an impossibility.

## 4.3. Rules of observation and refreshment

The 1861 to 1901 samples will inevitably be a product of the underlying base 1851 sample and the *rules of observation* and *refreshment* that determine who is included in the main VPS sample from 1861 onwards. Rules of observation relate to which persons included in the previous census should be included (observed) at the following census. Rules of refreshment relate to persons who were not observed in the previous census but who need to be entered into the VPS sample at the next census to maintain its cross-sectional representativeness of the changing British population. For reasons that will become clear below, both observation and refreshment rules are designed to keep the main VPS sample broadly representative of the British population over the 1851 to 1901 period.

---

[39] See the various contributions to *Annales de démographie historique*, 1 (2004) which is devoted to the TRA project. Two other historical projects, the *Geneva Database* and the *Historical demography of the Liege region* project, also have used surnames as a means of sampling, both selecting records where the surnames starts with the letter B.

[40] 'Nuptiality, Fertility and Mortality In Scotland, 1855-2001: A Family Reconstitution Project' draft project proposal supplied by Dr Peter Razzell.

### *4.3.1. Suggested rules of observation*

In order to facilitate comparison between the two, a strong argument can be made to model the VPS rules to determine who is observed on those of the BHPS.[41] This is also seen as the most effective means of keeping the VPS broadly representative of the British population between 1851 and 1901, with statistical weighting (as described in section 4.4.3) used to correct for any biases that might arise.

The basic structure of the main VPS sample is represented in Figure 4.1. This identifies three types of sample member: *original*, *additional* and *temporary*. These are defined as follows:

- **OSM** *Original sample member* – Everyone selected in the main 1851 VPS sample. Whether the 2% Anderson sample is used or a fresh sample drawn from a complete 1851 database, the sample should be household based.
- **ASM** *Additional sample member* – Sample member selected by refreshment rules.
- **TSM** *Temporary sample members* – Anyone who is not an OSM or COSM but lives from 1861 onwards in the same household of an OSM or COSM. An example would be a servant present in a household of an OSM in 1861 but no longer present in 1871.

These also give rise to two additional sub-types:

- **COSM** *Child of original sample member (and thence child of COSM)* – All children of OSMs born after the 1851 census (children of OSMs born prior to the 1851 census and strill in the household of the OSM are defined as OSMs along with their parents). Sequentially, all children of COSMs will themselves be defined as COSMs. Births will be traced via linkage with civil registration and subsequent census data.
- **CASM** *Child of additional sample member (and thence child of CASM)* – Child of sample member selected by refreshment rules (all children of ASMs will be CASMs, and all ). All children of CASMs will be themselves defined as CASMs.

Using these sample member definitions as a basis, three rules of observation can then be implemented, as follows.

[41] See http//www.iser.essex.ac.uk/ulsc/bhps/doc/vola/smfa.php#sampfoll.

*Observation Rule 1*

Attempted record linkage for all OSMs and COSMs is to be carried out from 1851 through to 1901. This linkage is to both Census and Civil Registration data (to link both marriage and death events). Where linkage cannot be made and a person might still be alive at a later census year linkage to all census and Civil registration years at which the person might be alive (e.g. aged < 100) will be attempted.

*Implications:* All successfully linked OSMs and COSMs will remain in observation, i.e. remain in the main VPS sample. Ceasing of observation is either due to known death (record linkage to Civil Registration death data), or an indistinguishable mixture of emigration and attrition caused by failed linkage to either census data (i.e. OSM or COSM still alive) or to civil registration death data (i.e. OSM or COSM dead).

*Observation Rule 2*

Attempted linkage of all ASMs and CASMs is to be carried out from their first census of selection onwards to 1901. This linkage is to both Census and Civil Registration data.

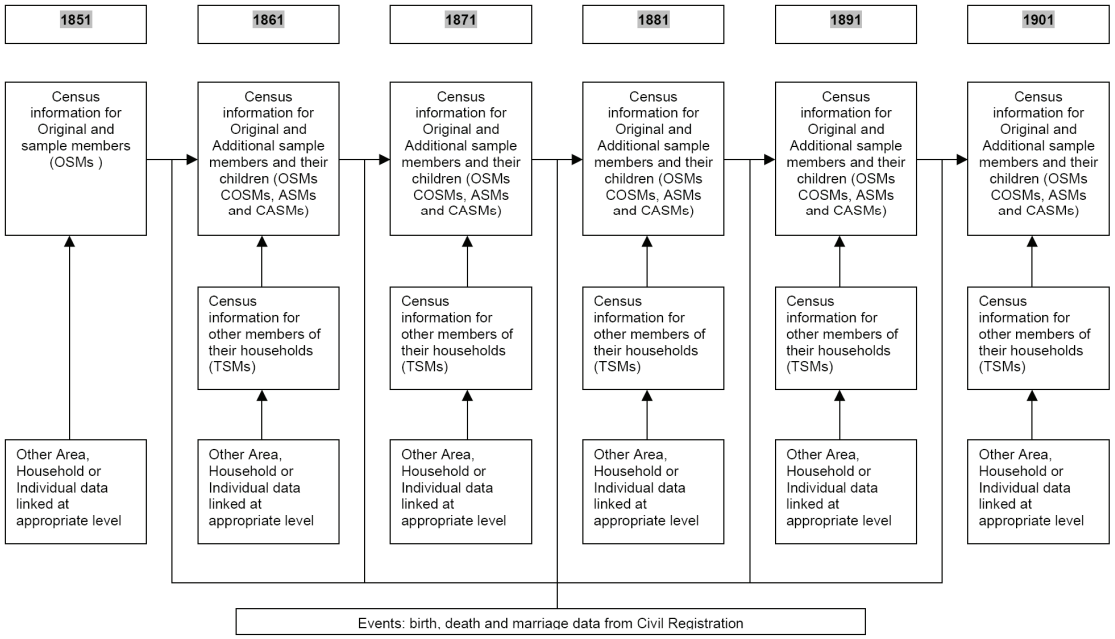
*Implications:* All linked ASMs and CASMs remain in observation, i.e. remain in the VPS sample from the year they are selected (via the application of refreshment rules). As for OSMs and COSMs, where linkage cannot be made and a person might still be alive at a later census year, linkage to all census and Civil Registration years at which the person might be alive (e.g. aged < 100) will be attempted. Ceasing of observation is either due to known death (record linkage from civil registration death data), or an indistinguishable mixture of emigration or attrition caused by failed linkage to either census data (i.e. ASM or CASM still alive) or to civil registration death data (i.e. ASM or CASM dead).

*Observation Rule 3*

Attempted linkage of TSMs is to be carried out from their first year of inclusion. This linkage is to both Census and Civil Registration data. TSMs only present in one Census year will not have attempted record linkage to Civil Registration data for the following decade.

*Implications:* TSMs are only maintained in the main VPS sample if they continued to reside in the same household as one or more OSMs, COSMs, ASMs or CASMs. Other reasons for ceasing of observation are death, emigration or failed record linkage as described for OSMs and ASMs.

Fig. 4.1: Proposed structure of the VPS



### *4.3.2. Immigration and rules of refreshment*

Refreshment of the sample via additional sample members (ASMs) is essential to keep the VPS representative of the growing and changing immigrant population over the 1851 to 1901 period. There are two possible methods to deal with immigration. These are outlined below, although only the second is deemed of suitable merit for a fully linked VPS due to uncorrectable biases in using this method.

*Method I*

This is to only make use of information directly observed by the census, this being the total overseas-born population at each census year. This figure is then used to generate the number of additional sample members (ASMs) that need to be added to the VPS at each census year, and then sample from these at the appropriate fraction.
*Advantages:* This method requires no additional resources for the full VPS.
*Disadvantages:* While the main VPS sample would be representative in terms of absolute numbers of overseas born at each census year, there would be a selection bias towards emigrants present in earlier census years and against immigrants the later they arrived in Britain. This bias is deemed unacceptable.

*Method II*

Starting at 1901 and working back to 1851 (if a full database is available) or 1861 (if not), attempt to back-link all persons recorded as being born overseas to obtain the earliest census at which they were observed. Allowing for some contamination due to failed record linkage (when the immigrant was in fact in Britain at an earlier census year but could not be traced back to it), this method provides both total numbers and the individual identities of persons arriving in Britain between each census.
One can use this information to randomly sample ASMs from *new* immigrants at the appropriate sampling faction and thereby keep the main VPS sample representative of the changing immigrant population over the 1851 to 1901 period.
*Advantages:* This method allows the greatest possible accuracy. The only source of appreciable bias is if failed back-linkage is relatively common and not at random with respect to immigrant characteristics.
*Disadvantages:* The time/expense of attempting to back-link all overseas born people could be considerable.

#### 4.3.2.1. Hypothetical example illustrating the two methods

Let us assume that complete record linkage of individuals can be obtained. For arithmetical simplicity let us also assume the following characteristics of a hypothetical British population in 1851.

| | |
|---|---|
| British born population | 10,000,000 |
| Overseas born (immigrant) population | 100,000 |
| Total | 10,100,000 |

Then assume both British-born and immigrant populations increase by 10 per cent by 1861 due to fertility (ignoring mortality to keep the arithmetic simple). Then assume a further 50,000 immigrants arrived between 1851 and 1861 (and ignoring the subsequent increase via fertility by 1861 to keep the arithmetic simple). This would give the following population profile in 1861.

| | |
|---|---|
| British born population present in 1851, still present 1861 | 10,000,000 |
| Births to British born population between 1851 & 1861 | 1,000,000 |
| Births to overseas born population present in 1851, still present 1861 | 10,000 |
| *Total British born population* | *11,010,000* |
| Overseas born population present in 1851, still present 1861 | 100,000 |
| Overseas born population having immigrated between 1851 & 1861 | 50,000 |
| *Total overseas born population* | *150,000* |
| *Total* | *11,160,000* |

Now assume that the main VPS sample constitutes a 2 per cent sample of the 1851 population and is representative of both native and overseas born in 1851. Since the hypothetical 1851 population is 10.1 million, this would mean that there would be approximately 200,000 British born original sample members (OSMs) and 2,000 overseas born OSMs.

Assuming this 1851 sample increases in line with the total population (i.e. a 10 per cent increase by 1861) and all persons are linkable between 1851 and 1861, by 1861 one would have the following VPS sample (prior to applying refreshment rules):

| *Hypothetical VPS sample by 1861* | |
|---|---|
| British born population present in 1851, still present 1861 | 200,000 |
| Births to British born population between 1851 & 1861 | 20,000 |
| Births to overseas born population present in 1851, still present 1861 | 200 |
| *Total British born population* | *220,200* |
| Overseas born population present in 1851, still present 1861 | 2,000 |
| *Total overseas born population* | *2,000* |
| *Total* | *222,200* |

This would no longer be representative of the immigrant population, since none of the 50,000 immigrants arriving between 1851 and 1861 would be in the sample. In 1861 the percentage of immigrants in the population as a whole would have been 1.3 per cent, while in the 1861 VPS sample prior to applying refreshment rules it would only be 0.9 per cent.

#### 4.3.2.2. Applying refreshment rules under Methods I and II

From the census figures one can directly calculate the number of overseas born in 1851 and 1861, from which one can derive the total number of immigrants arrived between 1851 and 1861. Via an appropriate sampling method, one could select 2 per cent of this number from all immigrants recorded at 1861 to be included as additional sample members (ASMs). Thus, to return to the hypothetical example given above, if it was estimated that 50,000 immigrants arrived between 1851 and 1861, it would be possible to randomly select as ASMs 1,000 (2% of 50,000) overseas born from the 1861 census. However, in doing this, it would not be possible to determine when these immigrants arrived in the country, thus the chances are that a relatively high proportion of them may have arrived before 1851. Assuming that two-thirds of the selected ASMs in fact arrived prior to 1851, this would have the following hypothetical effect, as shown below.

Whilst the British born element of the VPS sample would remain at 2 per cent by 1861, due to the inability to differentiate between overseas-born who arrived before or after 1851, the next effect of Method I in this hypothetical example would lead to an over-sampling of immigrants who arrived before 1851 (2.67% instead of 2%) and a substantial under-sampling of those who arrived after (0.67% instead of 2%). Since the magnitude of the bias would be unknown if using Method I, it could not be corrected by weighting. If Method I were followed, this bias would increase as time went on to 1901, such that immigrants arriving between 1891 and 1901 would have the greatest bias to-

wards under-representation, while the immigrants arriving prior to 1851 would continue to be the most over-represented group.

| *1861 VPS following Method I refreshment rules* | |
|---|---|
| British born population present in 1851 and still present in 1861 | 200,000 |
| Births to British born population between 1851 & 1861 | 20,000 |
| Births to overseas born population present in 1851 and still present in 1861 | 200 |
| *Total British born population* | *220,200* |
| Overseas born population (OSMs) present in 1851and still present in 1861 | 2,000 |
| Overseas born population (ASMs) having arrived by 1851 | 667 |
| Overseas born population (ASMs) having arrived between 1851 & 1861 | 333 |
| *Total overseas born population* | *3,000* |
| Total | 223,200 |

In comparison, since we are assuming that all possible record linkage is made, whilst more time consuming and arduous the rules proposed for Method II ensure that the outcome is entirely accurate, with as near as possible representative 2 per cent sample of those immigrating between and census years. Thus, it is clear that Method II provides the more acceptable refreshment strategy. However, since this carries substantial resource implications, it is important to determine what these might be. This is detailed in the following section.

Furthermore, even with real data the second method will not be perfect, since record linkage will most probably not be 100 per cent. This means that some immigrants who were present at a previous census will not be matched, and thereby added to the group of assumed new immigrants since the previous census. This group will then comprise all newly arrived immigrants (barring those not recorded in the census) as well as some contamination with those arrived earlier (but could not be back linked). The net result will be for the fully linked VPS sample to contain the correct proportion of newly arrived immigrants at each census year, but to over-represent longer established immigrants. The magnitude of this bias will be directly proportional to the frequency of failed back linkage. This frequency will not be estimable, but use of historical estimates of the total number of immigrants arrived at various time periods will allow an indirect estimate of the likely extent of the problem in the VPS sample.

Based on this exploration of the possible methods, it is now possible to pose various rules for refreshment, as follows.

*Refreshment Rule 1. ASMs to compensate for immigration*

The new immigrant population (those arrived since the previous census) will be sampled at each census (1861 to 1901) to match the VPS sampling fraction (the rest of the population) for each census year.

For example, if the VPS sample at 1871, prior to the addition of ASMs, comprised a 2% sample of the population that had not immigrated since 1861, then 2% of the immigrant population arrived since 1861 will need to be sampled, making the resulting VPS 1871 sample a representative 2% sample of the total British population in 1871.

*Rationale*

Estimation of the population that has immigrated since the previous census is practical but resource intensive, requiring full backward linkage (to census and civil registration data) of all overseas born persons for each census year. In this way, one can arrive at estimates of both absolute numbers and individual identity of persons who have immigrated since the previous census, and hence a sub-population from which to sample ASMs to compensate for immigration since the previous census.

The need for the VPS to attempt record linkage is detailed in 5.3. This illustrates why this is the only viable method for keeping the VPS sample representative of the immigrant population over the 1851 to 1901 period. Due to record linkage failure the VPS will still contain some bias towards over-representation of long-standing immigrants but should contain the correct proportion of newly arrived immigrants (see section 4.3.2. for details).

Note that since all immigrants will need to be traced over all census years, this makes an immigrant booster sample possible at little extra cost. This would provide a valuable resource for researchers interested in the differential assimilation of the various immigrant communities. This potential is expanded on in section 4.3.6.

### *4.3.3. Rules of exit/ceasing to be observed*

These are entirely defined by the preceding rules of observation and refreshment, but for ease of understanding they are also presented separately below. All persons who cease to be observed will be given a variable indicating the reason for this.

For *all* types of VPS sample members exit will occur due to:

- Death (confirmed by linkage to Civil Registration death data)
- Emigration
- Attrition (failed linkage to census data, i.e. person still alive)
- Attrition (failed linkage to Civil Registration death data, i.e. person dead)

The two sub-categories of Emigration are listed since they will not be distinguishable at the individual level in the VPS.

For TSMs exit will additionally occur when:

- The TSM is no longer resident in the same household as one or more OSMs, COSMs, ASMs or CASMs

### *4.3.4. Additional refreshment rules*

In the rules presented in section 4.3.2 refreshment was only carried out to maintain the representativeness of the VPS sample in terms of the immigrant population. In theory, additional sample members (ASMs) could be used to correct any other emerging bias in the VPS sample, in particular for loss of sample size due to attrition, or to adjust the VPS to make its profile more representative of the British population. It is possible, therefore, to consider also the following rules of refreshment.

*ASMs to compensate for attrition*

People added to make up (cross-sectional) sample size in the latter census years if failed record linkage leads to excessive loss of cross-sectional sample size.

*ASMs to make post-1851 VPS samples more representative of the British population*

The post 1851 VPS samples might diverge in aggregate characteristics from the population as a whole (in terms of characteristics measured in the census).

Selecting ASMs to boost sample size in the latter years of the VPS, particularly if this is accompanied by back linkage as outlined below, will increase the bias towards inclusion of persons who can be linked over time. To the extent that record linkage failure does not occur at random with respect to socio-economic and demographic characteristics, adding such ASMs will increase the bias of the main VPS sample towards characteristics that are unrelated to attrition. Taken to its extreme, one could (using back linkage) select an extra sample that was not subject to any attrition (i.e. select only individuals who could be traced back until 1851 or birth), but one could not hope to address the biases involved in such a sample via attrition weights.

The second type of ASM, that to adjust the sample to make it more representative of the British population at a given census year, is an issue that can be more effectively addressed via post-stratification weighting (see section 4.4.4).

It is therefore not recommend that the VPS uses either of these rules of refreshment as the former makes the problem of attrition bias worse, and the latter is better catered for using post-stratification weights.

### *4.3.5. Back-linkage of additional sample members (ASMs)*

Refreshment in an historical longitudinal survey raises the interesting possibility of 'back-linkage' of additional sample members (ASMs) back to 1861 (or 1851 if a full database is available). Such back linkage would maximise the longitudinal value of these records and would give the largest possible 'constant panel' in the VPS. However, this would mean that the exact sample design of 1851 would be impossible to specify, since it would no longer be defined by a set of sampling rules, rendering design effects unknown, and thereby uncontrollable for statistical analysis. Additionally, back linkage could make earlier years unrepresentative of the British population, so while ASMs might be added to make a latter year more representative of the British population (i.e. type 2 ASM in the previous section), this addition might make an earlier year less representative when these ASMs are back-linked, leading to an iterative sampling complexity that could not be handled by conventional software, making selection of such ASMs a highly complex mathematical task. For the reasons given, it is not recommended that the VPS should back-link ASMs. Indeed, the only recommended ASMs are new immigrants, who by definition cannot be back-linked in any case.

### *4.3.6. Additional immigrant samples*

Since complete back linkage of all overseas born persons is the main method to ensure the VPS is representative of the immigrant population over the 1851 to 1901 period, should this be implemented, it raises the attractive possibility of creating separate immigrant sub-samples. These might be either a representative sample of immigrants arriving in a given time period, or a sample of people from a specific geographical origin of particular interest (e.g. Irish). If the complete back linkage is undertaken to ensure representativeness of the immigrant population, then drawing additional immigrant 'booster' samples will have only modest additional cost and time implications for the VPS, and is thus an attractive proposition.

### *4.3.7. National and regional samples*

It is clear that some researchers will have interests that focus on Scotland, England, Wales, or England and Wales. The main VPS sample, based on the British population must be made so it can be used to analyse change within a constituent country via selection of cases in those countries and application of appropriate weights.

The research community may also be interested in the regional picture. Here the main VPS sample size becomes an important consideration. While regional weights could be created, the precision of resulting estimates may be suffi-

ciently low that results will be difficult to interpret. One key question that needs to be addressed is should the main VPS be designed to allow regional analyses, and if so at what level?

The initial cross-sectional sample design for 1851 will be an important consideration here. A geographically clustered sample design, which is typically used by contemporary surveys to minimise travel costs and interview times, would have the benefit of providing local pockets of well-sampled areas, allowing regional and local historians to focus on what might be called 'regional' profiles. However, the geographical clustering will be diluted over time, especially in areas of high in-migration. It may, therefore, be decided that for those interested in local and regional history will be better served by the possibility of additional discreet VPS samples that track the population of specific area over time (see below).

#### *4.3.8. Geographically delineated communities*

Among local and regional historians, there will be an interest in tracking changes in a given area over time. Even if the main 1851 VPS sample is geographically clustered, as mentioned already, clustering will be diluted over time due to migration. This will make the main VPS sample less useful for studying geographically delineated communities over the 1851 to 1901 period, particularly in areas of high in-migration.

To serve the local and regional historian research community, it might be desirable to include one or more additional samples that are explicitly geographically delineated. In these additional samples the rule of observation and refreshment might simply be residence in the area at a given census year, with destination noted for those who exit observation due to migration, and origin and other prior characteristics noted for those who enter observation due to in-migration. Indeed, most historical studies that have constructed a longitudinal dataset using census materials in the UK have tended to be geographically delineated. The key questions are how large such areas should be, and which areas should be selected, which poses both methodological and financial questions for the VPS project to consider.

## 4.4. Weighting and Sample Design Effects

#### *4.4.1. Background*

This section concerns the sample design effects, particularly the creation of weighting variables, in conjunction with the main VPS sample. The other possible supplementary samples mentioned previously (geographically delineated

samples and immigrant population samples) are unlikely to require weighting or raise other design effect issues.

The sampling strategy of the VPS detailed above has been designed to maximise the representativeness of the main VPS sample to the British population. Due to the nature of the data, some degree of bias is inevitable. Many sources of sample bias in the main VPS sample can be addressed by weighting, which will make the parameter estimates generated by weighted analysis *unbiased* (albeit generally with an accompanying loss of *precision*).

To understand more fully what weighted analysis entails, one needs to distinguish the three primary types of weight that can exist in any historical or contemporary survey dataset. These are:

- sample design or probability weights;
- non-response or attrition weights;
- post-stratification weights.

### *4.4.2. Sample design/probability weights*

Sample design or probability weights correct for cases (in this instance persons or households) having unequal probabilities of selection that result from sample design. It is important to note that non-equal selection probabilities can also occur due to differentials in response, which is corrected by non-response weights described in the following section. Minor discrepancies may also require adjustment if the sampling frame does not entirely reflect the population, and these would constitute a type of post-stratification weight outlined in section 4.4.4.

To illustrate how a sample design weight is calculated, consider a survey design that selects only one adult per household. Provided information concerning the number of adults per household is also enumerated, one can subsequently calculate sample design weights that correct for the lower selection probabilities of adults in multi-adult households. The general formula for a sample design weight is arithmetically very simple, it is 1 divided by the probability of selection due to the survey design. However, these are usually scaled, so the actual weight variable is proportional to this number. For example, if there are three adults in a given household the resulting sample design weight for the single interviewed adult will be proportional to 1/(0.33), i.e. proportional to 3. In a one adult household, the weight will be proportional to 1/1, i.e. proportional to one. In other words the influence of the former person is being increased threefold relative to the influence of the latter respondent to exactly compensate for the fact the former respondent is three times less likely to have been included in the sample.

As detailed below, whether sample design weights play a part in the VPS will depend upon the method of drawing the 1851 sample, and in particular is

the existing Anderson 1851 sample is used, or a new random sample is drawn from a complete machine-readable version of the census.

### *4.4.3. Non-response or attrition weights*

Before addressing the issue of attrition with respect to the VPS, it is instructive to introduce the main two types of non response that affect contemporary surveys: item and unit non response. Response rate refers to 'unit non-response', meaning that someone refuses to take part in the survey at all, in contrast to 'item non-response', which relates to refusing to answer specific questions. In contemporary surveys item non response is typically addressed via missing data methods, while unit non-response is addressed via attrition weights. As discussed at the end of this section the two phenomena are potentially more closely linked in the VPS.

In contemporary social surveys non-response weights compensate for differential response rates. They are typically obtained by defining *weighting classes*, which are based on information available for both responding and non-responding individuals or households. Such information typically relates to geographical location, primary sampling unit (PSU) characteristics (which are derived from other data sources, often the Census) and household and dwelling type or estimated age and sex (if a personal interview is attempted). Respondents in each weighting class are weighted to compensate for the proportion of non-respondents in that class. More formally, the non-response rate weight is proportional to 1 divided by the response rate for the weighting class, i.e. directly analogous to sample design weights. The utility of non-response weights is generally dependent upon the amount of information known about non-respondents.

In contemporary longitudinal surveys such as the BHPS and Birth Cohort studies, non-response rates tend to increase over time, since some people become fed up with repeated surveying or cannot be traced when they move. This leads to *attrition,* a progressive reduction of the sample size since the first survey. Since one knows detailed information about a former respondent who subsequently refuses to take part in the survey, one can create more detailed weighting classes than is usually possible for cross-sectional surveys, hence attrition weights generally correct for bias more effectively than non-response weights. The assumption of both non-response and attrition weights is that the characteristics of respondents and non-respondents within each weighting class are the same; only if they are will the weight be entirely accurate. The more information available, the more sensitive the weighting classes can be and the more accurate the weighting variable will be in correcting bias. It is also important to realise that the bias reduction of non-response/attrition weights will be variable specific (i.e. vary according to the characteristics one is examining).

Bias correction is at its highest where the characteristics under investigation are either entirely uncorrelated with non-response/attrition (i.e. no bias to correct) or are highly correlated with the variance of non-response/attrition explained by the weighting classes. Conversely, bias correction will be its lowest where the characteristics under investigation are highly correlated with non-response/attrition *and* uncorrelated with the variance of non-response/attrition explained by the weighting classes.

#### 4.4.3.1. 'Non-response' and attrition in the VPS

Attrition will be a major issue, indeed the major methodological issue, for the VPS to confront. Although attrition in the VPS is methodologically speaking analogous to that in contemporary longitudinal surveys, the causes of attrition are different. In the VPS, attrition will be caused by failed record linkage (between censuses) rather than cumulative non-response.

The rate of attrition due to failed record linkage will be the greatest single determinant of the representativeness of the VPS sample (since, as noted, attrition weights cannot entirely correct for attrition bias, especially given the relatively limited amount of information with which to construct weighting classes). The precise details of how attrition bias is likely to affect the main VPS are given in section 4.5.1.

#### 4.4.3.2. 'Unit' and 'Item' non-response in the VPS

As noted previously, unit and item non-response are often treated as separate issues in contemporary surveys, with only the former being addressed via weighting.

In the VPS, the analogous issue to item non-response is partially illegible or incomplete enumerators' returns. The completeness of information for an individual, most especially of forename and surname, is likely to impact on the success of record linkage and hence the likelihood of attrition. In this way item and unit non-response are linked. It is therefore important to examine whether item non-response, particularly of name information, is missing or incomplete at random with respect to other observed characteristics. If it is random, the risk that item and unit non-response are linked in a way that will affect subsequent analysis (other than that of personal names) is greatly reduced. Details on this issue are given in section 0.

### *4.4.4. Post-stratification weights*

Post-stratification weights (also known as population or calibration weights) are typically constructed after the other types of weights.[42] They are applied to

[42] Though calculation of non-response and post-stratification weights can be performed simultaneously.

make the data even more representative of the population, or more specifically certain observed characteristics of the population (termed auxiliary or calibration variables). Not only do post-stratification weights allow for more accurate population total estimates, they typically reduce non-response/attrition bias further (over and above the application of non-response/attrition weights) and improve precision of estimates (unlike survey design and non-response/attrition weights which generally reduce precision). For all these reasons, post-stratification weights are generally beneficial.

In contemporary social surveys, information on the population is usually derived from the decennial Census of Population, and this will be the case for the VPS, with full British population-level information available for all census years from 1851 to 1901. Whereas probability/sample design and non-response/attrition weights result from a simple computation (1/selection probability), the calculation of post-stratification weights is mathematically complex, requiring iterative algorithms (the Newton-Raphson method) that maximise the fit of the data to the calibration/auxiliary variables of the population. This procedure is called 'raking', and requires specialist software. The raking procedure iterates until the data best match all control variables, and computes the post-stratification weight(s) accordingly. The Office for National Statistics currently use a SAS-based macro called CALMAR to calculate post-stratification weights,[43] and the BHPS is using CALMAR from the release of Wave 14 onwards. Consequently, the VPS team acquired and trained themselves in the use of CALMAR in order to estimate its appropriateness for the VPS. An introduction to the CALMAR software and the alternatives is given in section 3.1.1.

## 4.5. Recommendations on weighting

Section 4.4. introduced the three main types of weighting variable. This section examines the detailed implications of each type with respect to the VPS. The following section then synthesises this information to provide guidance and recommendation for the weighting strategies for the main VPS.

Sample design weights are only pertinent to the initial 1851 sample as this is the only sample for which design can be specified. The method of selecting this sample will be governed by data availability, in particular whether a full transcription of 1851 is available to sample from, or whether the 2 per cent Anderson sample is the only source of an initial sample. If a full transcription is available, it is recommended that the main VPS sample for 1851 should be made using made using an equal probability of selection sampling method (EPSEM) as this would negate the need for sample design weighting, or any

[43] See http://www.ccsr.ac.uk/esds/events/2004-03-12/documents/barton.ppt. The ONS may switch to Statistics Canada's Generalized Estimation System (GES) programme in the future.

further sample design issues (clustering and/or stratification effects). If the Anderson sample is used, clustering effects, and how to adjust for them using commonly available software, would need to be explained for secondary analysts in the VPS data documentation.

### *4.5.1. Non-response/attrition weights and the VPS*

As noted previously, bias caused by attrition due to record linkage failure not being random with respect to socio-economic and demographic characteristics is likely to be substantial, and so weighting to reduce such bias is essential.

*A priori*, record linkage failure may be thought to be linked to literacy and numeracy (via the ability to spell out one's name and give a comparable age at each census), and migration behaviour (those who move most often and furthest being less likely to be linked), and thereby (or possibly independently) also age. Any such relationships would introduce important biases in the VPS sample.

Taking into consideration these issues, the methodology recommended for the VPS is as follows:

1) Identification of directly enumerated and subsequently derived individual-level characteristics that are correlates of attrition, using similar methods to the samples analysed by this project;
2) Creation of weighting classes based on (1) above;
3) Calculation of attrition weights based on (2) above;
4) Consideration of methods other than weighting for dealing with attrition bias;
5) Consideration of residual bias that might not be adequately controlled by (3) and how to identify and deal with it.

### *4.5.2. Post stratification weights and the VPS*

Post-stratification is logically reasonably simple but computationally intensive. For selected population level characteristics (calibration/auxiliary variables), iterative raking will provide the weights that maximise the fit of the sample data to the population for these characteristics.

The project identified and evaluated three specialist software packages that calculate post-stratification weights. These are summarised as follows:

- CALMAR – a freely available SAS macro developed and distributed by Instititut National de la Statistique et des Etudes Economiques (INSEE), France.
- GES – a SAS based programme developed by Statistics Canada and requiring $ (Can.) 30,000 per site license
- G-Calib – an SPSS based programme developed by Statistics Belgium.

This exercise suggested that the CALMAR warranted further investigation, especially as it would most probably be made available to the VPS at no cost. It was thus acquired for testing with a small selection of sample data. Self-training in the package was necessary to ensure it was reasonably simple to learn with the available documentation (which is only partly in English). It was reasonably simple to use and worked examples were followed using both the supplied examples and some BHPS data.[44]

As a result of this exercise, the use of CALMAR for post-stratification weighting by the VPS is recommended, barring any subsequent developments in this field. To calculate post-stratification weights using CALMAR the VPS will need to firstly, identify auxiliary/calibration variables if complete multi-way stratification is not possible, then calculate post-stratification weights at individual and household level with the CALMAR package using the variables selected.

## 4.6. Likely weighting variables that the VPS will need to create

It is important to consider that there will be several dozen weighting variables in the final VPS, even assuming that attrition and post-stratification effects are combined into a single variable. At minimum, the following variables are anticipated:

- Cross-sectional attrition *and* post-stratification weights for each census year (i.e. 6 weights)
- 'National' weights for England, Wales and Scotland (cross sectional only, for each of the censuses 1851 to 1901, i.e. 18 weights)
- Full set of longitudinal census weights from 18$x$1 to 18($x$+1)1 or 1901 (15 weights)

Even by combining attrition and post-stratification weights, this gives a total of 39 weight variables, and this can be doubled if both person and household level weights are created.

The final VPS should also consider whether to scale up weights to provide a suite of cross-sectional *grossing* variables – to allow the secondary analyst easy generation of national level estimates of actual numbers of people or events in Great Britain (or one of its constituent countries) at one of the census years. This would seem desirable provided clear user guidance was given to avoid using such variables to examine standard errors in SPSS (outside the Complex Samples module) because of the erroneous precision that would occur. There would be six potential grossing variables: cross-sectional attrition *and* post-stratification grossing weights for each census year. Since country will be indi-

[44] This was undertaken by Alasdair Crockett in collaboration with Annette Jackle of ISER, University of Essex.

cated in the database, a single grossing variable will permit national estimates for the constituent countries of Great Britain if so desired.

## 4.7. Data dissemination issues

The previous section has dealt with weighting issues from the viewpoint of creating the full VPS. The function of weights is to allow the secondary analyst to reduce bias of their estimates and to generate estimates of appropriate precision. To ensure this is achieved, the VPS will have to do more than simply compute the necessary range of weighting variables, it will need to provide clear documentation and advice on statistical packages that can use weight variables appropriately, especially given the multiple weights that will be needed. This is of particular importance as these issues are often ill understood by contemporary social scientists and are generally less likely to be well understood by historians. Lack of complete information about weighting variables is a major deficiency in some large contemporary social surveys. In particular, four deficiencies are frequently observed:

- paucity of documentation relating to how weights were calculated;
- paucity of guidance about what weight variables should be applied according to the research questions being addressed by the secondary analyst;
- weight variables are often multiplied into a single variable, so that constituent effects cannot be investigated by the secondary analyst, particularly as the weighting classes and population totals used for post-stratification weights are typically not known to the secondary analyst;
- lack of guidance on what software packages should be used to conduct weighted analysis, particularly the problems with SPSS not adjusting precision correctly in weighted analysis (unless the Complex Samples module is used) and the inability of relational databases, often preferred over statistical packages by historians, to handle weights at all.[45]

As such, in creating the final user version of the VPS, a particularly important task will be to generate clear documentation and guidance to users on the application of these weights.

[45] Prior to the version 12 Complex Samples module, SPSS did not adjust precision according to the effective sample size, but rather used the scale of the weighting variable to determine the sample size, such that weighting variables with a mean of > 1, and in particular grossing variables, will lead SPSS to generate results with far too high a precision. The other commonly used packages of STATA and SAS do not suffer from this defect. SPSS 13 is introducing greater functionality, but this will still be limited to s separate module, meaning that naïve users, or those at institutions not paying for the module, will continue to generate results of incorrect precision when carrying out weighted analysis in SPSS.

# 5. Record-linkage strategies

## 5.1. Approaches to record linkage

Despite the fact that it is often embarked upon, record linkage is undoubtedly problematic. One of the problems of coming to terms with the theory of record linkage is that although is a technique which is applied in several disciplines, it is known by different terms and uses a non-standard set of terminologies, with little correspondence across the range of disciplines in which it is applied.[46] Indeed, one of the curious aspects of historical record linkage is that the existing literature makes scant and more usually non-existent reference to developments in other fields.

Basically, there are just two general approaches that have been adopted: linking techniques (both deterministic and probabilistic) and match merging techniques. The latter are generally easier to undertake but as a result are prone to inaccuracies and incomplete matches due to data errors and omissions. In contrast, both forms in the first group tend to be more complex but result in more accurate and complete matches.

### *5.1.1. Match-Merge Methods*

Match merging is used in a lot of data processing tasks, and essentially is undertaken by employing keys or fields within variables on each file being linked or merged to produce the match, where records from two of more files are combined when the respective keys from each file are the same. An example of a match merge would be when, say records from different tables in an Access database are joined (via a defined Relationship) using a identifier that is common to both tables, such as unique employee number or customer reference. Such a technique can be easy and fast, but will lead to errors is the key being used is incomplete for all records or is error prone. In the case of historical sources simple match-merging is not realistic since very few historical sources have appropriate personal identifiers.

[46] For example, although demographers and epidemiologists refer to record linkage the same process is know as entity identification, object isomerism, instance identification, entity reconciliation, list washing and merge/purge by statisticians, computer scientists and marketing experts. See, Gu *et al* 'Record linkage'.

### *5.1.2. Deterministic Linking*

Rather than overly relying on a (hopefully) unique personal identifier, deterministic record linking seeks to overcome the limitations of match merging by comparing identifying information from each of two files and assigning points or scores for each agreement. Usually applying multiple criteria, only records with a total score over a predefined threshold are linked. Indeed, it is this method that forms the basis of nearly all record linkage that has been undertaken using historical sources. A hypothetical deterministic link might score two records using the following criteria:

- 20 points for a complete agreement on surname;
- 10 points for a complete agreement on forename;
- 5 points for an agreement on an acceptable surname variant;
- 5 points for an agreement on an acceptable forename variant;
- 5 points for a complete agreement on birth year;
- 2 points for a birth year within +/- three years;
- -10 points if gender birthplace does not agree.

Higher points reflect higher importance of the criterion. Applying this hypothetical procedure might deem that linked records with scores of 30 points or more are acceptable. Unlike match/merging deterministic, linking will often result in unresolved or ambiguous multiple links. Often three states of linkage pairs emerge:

- 'True' links (defined by the lack of alternatives or the high value of their score);
- 'False' or 'rejected' links (defined by the failure to link or the low value of their score);
- 'Ambiguous' links (defined as those were two or more possibilities arise and scores are tied or not significantly different).

Indeed, in most record linkage exercises using historical sources, automated techniques are used only to resolve the first two types of linkage pairs, with all ambiguous links being written out of the database in order to be resolved manually. Thus, most historical record linkage exercises are in reality only semi or partially automated. This is not realistically an option open to the VPS since the number of ambiguous links could be too high to perform manually. Moreover, there is an objection to this practice on methodological grounds given that manual linkage (based on intuition) may introduce a bias, especially if undertaken inconsistently by different researchers. Because of this concern, it has been suggested that if it is possible to produce consistent rules to decide

between ambiguous links, then it must be possible to specify these in programmable form.[47]

A major problem with deterministic linking arises from the difficulties of establishing appropriate scores for individual agreement criterion and setting an appropriate threshold for linking given that scores and thresholds cannot be set empirically. Usually in deterministic linkage, and very often in historical cases, the number of points awarded for an agreement between a pair of records is arbitrarily set, albeit through extensive trial and error and historical judgement on what constitutes a 'good' or 'likely' match . Returning to the example of scores given above, whilst it may appear obvious that complete agreement on a surname spelling should be more important than agreement on a variant spelling, it is not intuitive that it is exactly four times as important. Clearly, the relative weighting of the criteria as reflected in the assignment of points, is critical to the success of the linking process. Scores need to reflect the relative importance of an agreement. However, the relative importance of an identifier can vary from case to case. For example, in linking those at risk to marry in a census to entries in the civil registers, birth year will not be as important as in linking between censuses, due to the fact that since marriage registers apply to a restricted age cohort chance agreements on birth year are more likely. Another limitation is that deterministic linking does not provide a mechanism for scaling agreement points. For many identifiers, the relative importance of an agreement depends on the value. For example, comparing surnames, agreement on a rare name such as Schürer should receive a higher score than agreement on a relatively common name such as Jones, with a relatively uncommon name such as Crockett, somewhere between the two.

Lastly, a further problem in assigning relative scores in deterministic linking is with what fields of variables to include. There is a temptation – which most historians seem unable to resist – to use the maximum amount of information possible in assigning scores, with a correspondence on any two possible matched pairs receiving a score. But some attributes relating to an individual are at risk to change over time, whilst others are not. Thus, although some individuals will remain in the same occupation over their working life, others will not, and change is potentially more likely to occur at certain points in the life cycle than others. Thus, if agreement on occupation is used to add to the accumulative score then this could give rise to a biased outcome in favour of the occupationally immobile. Likewise, if points are awarded to agreement on place of residence then this could result in a bias toward the non-migratory elements of society. [48] Because of this problem, following the work of Ferrie,

[47] Schofield, 'Automatic family reconstitution' and Schofield, 'The standardization of surnames'.

[48] Although not a problem for the VPS, the most common problem of this sort in historical linkage exercises is bias introduced as a result of liking from multiple sources, where the sources in question record sections of the population disproportionately. See King, 'Record

Ruggles in his US census-based linking exercise has advocated the use only of fields which in theory should not change over an individual's life: surname, forename, birth year, place of birth.[49]

### *5.1.3. Probabilistic Linking*

Probabilistic linking is very similar to deterministic linking in that it uses multiple criteria and scores to establish record links. The main difference lies in the manner in which points and thresholds are set. With deterministic linking, agreement points and linkage thresholds are set outside of and prior to the linking process. In probabilistic linking agreement scores are determined by the data themselves and are scaled, relative to the value of the identifier. Thus the main weakness of deterministic linking, the arbitrary and rigid assignment of agreement scores, is overcome. Likewise, probabilistic linking also makes use of *disagreements* in linking records as well as *agreement*, a factor that is largely ignored in deterministic linking. However, whilst probabilistic record linking resolves some problems of both match merging and deterministic linking, it does so at the cost of complexity, and it is because of this problem that it has been applied rarely in historical studies.[50] It involves multiple, non-trivial steps to calculate weights, set linking thresholds, and link the data.

The available literature on probabilistic linking puts forward a number of approaches, where probabilities are calculated both directly and explicitly,[51] including expectation-maximisation,[52] and Bayesian models.[53] Detailed description of such technique can be found elsewhere, and it is not intended to repeat these here, other than to conclude that probabilistic linking offers some clear advantages to the creation of a VPS, principally for the reasons specified below.

First, a VPS will differ from most other previous historical linkage exercise, and especially those undertaken by family historians and genealogists, in one important respect, namely that the VPS will seek to link records in such a way that it minimizes bias in the underlying sample. This differs from most other linkage exercises using historical sources which aim essentially to maximize 'accuracy'. Thus, in resolving ambiguous or multiple links in the VPS it will not matter if a 'wrong' decision is made as long as the decision can be shown

---

linkage in a proto-industrial community' and King, 'Multiple-source record linkage' as an example.

[49] Ferrie, 'A new sample of males'. Ruggles, 'Linking historical censuses'.

[50] Although not strictly probabilistic linking, a related approach is provided in Harvey and green, 'Record linkage algorithms'. See also Harvey, Green and Corfield, 'Record linkage theory and practice'.

[51] See, for example, Corpas and Hilton, 'Record linkage', and Gu *et al*, 'Record linkage'.

[52] For example, see Jaro, 'Probabilistic linkage'.

[53] See Larsen and Rubin, 'Iterative automated record linkage' and Fortini, *et al*, 'On Bayesian record linkage'.

to have no impact of resulting bias. Probabilistic linking helps to overcome such a problem.

Second, very few, if any, historical record linkage exercises have attempted to measure the confidence that one has in making the link, in other words the probability that the link is 'correct'. Probability linking enables such calculations to be made, and as a result, the VPS would be able to attach to each linked pair a confidence value which researchers could use to filter or weight their subsequent analyses accordingly. This has a clear advantage over most record linking exercises in which, effectively, all successfully linked pairs are given the same weighting or value.

Third, whereas most previous historical record linkage exercises have essentially been local in character, usually taking census records for a given place and attempting to link them to census records of the same place at given points in time, the VPS will be on a national scale. By employing complete national datasets it will be able to utilize probabilities drawn both internally from the data and externally from related sources. This is best illustrated through a couple of examples. Given the VPS will have recourse to complete national census material, it will be possible, for example, to identify each possible surname and forename variant and the probabilities of these occurring in the datasets to be linked. This could be done both nationally and regionally, and can thus be used in resolving the likelihood or probability or one matched pair being 'correct' over another. Equally, combining both the census micro data and aggregate macro data (as recording in the published reports of the Registrars General and elsewhere) it would be possible to calculate the probability of vital events, such as death or marriage, occurring for any individual under observation between two time periods.[54] Once calculated, such probabilities can be used within the linking process to help resolve multiple links and 'dead' links. Likewise, probability linking can 'learn' from other related test linking exercises.[55] Thus the VPS could benefit from the experience gained from other appropriate linkage studies. An example would be the exhaustive linkage exercise of six Kentish parishes undertaken by Perkyns which measured the discrepancy in the recording of ages across censuses, both studying the extent of the discrepancy and its variance by a number of factors such as gender, age, social class and birthplace migration.[56]Again, such probabilities as the likelihood of an individual's age (birth year) being recorded differentially across census years 'learnt' from previous studies could be used to inform and guide the linking process.

---

[54] As a by-product of the project aggregate statistics on vital events nationally would be tabulated and used to estimate the expected numbers of individuals, by age, that one would expect to have died or emigrated between census years. Although complex, with complete national individual-level census data it is believed that it would be possible to calculate such probabilities at a regional as well as national level, extending the previous work of Baines, *Migration in a mature economy*.

[55] See, for example, Elfeky, Verykios and Elhagarmid, 'TAILOR'.

[56] Perkyns, 'Age checkability'.

However, whilst offering attractive solutions, probability linking does have some drawbacks. In addition to the complexity issue, one of the inherent logical requirements of probability linking is that pairs are matched only on criteria that exist for all possible pairs. In other words, it takes as input only those variables within the datasets that exist for all individuals. This could present a problem since historic census data contain only a small number of variables that are 'fixed' and which exist for each individual: gender, forename, surname, birth year, and birthplace. Information on occupation, marital condition and place of residence obviously also exist for each individual, but these are not fixed in the sense that they are at risk to change over an individual's life.[57] Thus, there is relatively little fixed information to make links from and in the case of the VPS, one might need to rely on non-fixed information and, moreover, incomplete information to resolve multiple pairings. For example, the names, when known, and birth years of parents, which would be available for some, but not all, of the population under observation.

Related to this problem is the fact that probability linking techniques are essentially aimed at pairing entities across two or more data files, where the entity in question – usually a person – is independent of all other entities. This is not true of the VPS which will aim to link individuals *and* households, where one is logically consistent with the other. This hierarchical issue is discussed in the following section.

In conclusion, whilst probability linking offers many attractions and advantages, for pragmatic reasons it is recommended that the VPS adopt a hybrid approach, combining aspects of probability linking with aspects of deterministic linking.

### *5.1.4. Households and individuals*

Despite the number of projects that have undertaken linking exercises between historic censuses, it is curious that very few would appear to take the household component into consideration in devising their methodologies.[58] Most of the literature based on census linking concentrates on individuals both from a methodological and analytical view point. Yet the VPS, as the discussion on sample design in section 4.2 makes clear, will be a panel of individuals within households, in which the household is an analytical unit as much as individuals. This has important implications for linking, since both individual and house-

[57] Disability is also sometimes recorded for individuals in the nineteenth-century censuses, but this can be inconsistently recorded.

[58] Exceptions include Mineau, Bean and Anderton, 'Description and evaluation of linkage', and Pouyez, Roy and Martin, 'The linkage of census name data'. The use of household information in linking also forms an important element in the project linking census and civil registration records for the Isle of Skye, Kilmarnock, Rothiemay and Ipswich, run by Dr E. Garrett, based at the Cambridge Group for the History of Population and Social Structure. See, Davies, Garrett and Reid, 'Nineteenth-century Scottish demography'.

hold units must be logically consistent. For example, hypothetically, if a married individual A in household X in one census is linked to household Y in a subsequent census, then one would expect A's spouse (assuming they are still married) to also be linked to household Y, even though an apparent 'stronger' link may occur between the spouse and a different household. This may appear to be obvious, but since most record linkage focuses on the individual it is possible for such logical inconsistencies to occur.

The household can also play another important role in record linkage in that although the number of fixed variables for any individual in an historic census is relatively small, when placed in the context of the household in which they are resident, the number of additional attribute variables associated with a given individual can increase markedly. For example, in the nineteenth century most individuals lived within families, within households. In the 1851 Anderson sample 30.4 per cent of individuals lived with a spouse, 37.9 per cent with a father, 41.5 per cent with a mother and 27.7 per cent with one or more children. Most significant of all is the fact that only 22 per cent of the population did not have either a co-resident spouse, parent or child. Thus, for 78 per cent of the population it is possible to assign attribute variables such as parent's names, birth years and birthplaces, spouse's name, birth year and birthplace, and a range of children's names, birth years and birthplaces. Consequently, for most individuals, in addition to the fixed variables, it is possible to assign up to three or four times as many associated attribute variables. Although such attribute variables run counter to the basic theory underlying probabilistic linking, since the information is not complete due to the fact that not all individuals being linked have parents, a spouse of children, they can be a significant aid in resolving ambiguous multiple links. Moreover, for the VPS it will be critical that such attribute variables are taken into consideration during the linking process in order that household linking remains logically consistent with individual linking.

## 5.2. Standardisation of data

### *5.2.1. Name standardisation*

Much of the literature on historical record linkage has concentrated on the problem of standardising names as recorded in the source materials – mostly surnames, but the same problem can be applied to forenames.[59] This is seen as

[59] The summary in this section is based on a fairly exhaustive review of the literature. For reasons of brevity, only key references are footnoted in the text, however, a full list of the materials consulted is provided in the bibliography to this publication. Of the texts listed there, a useful starting point for those wishing to research the subject is Snae and Diaz, 'An interface for mining genealogical nominal data', Lait and Randell, 'An assessment of name matching algorithms' and Nygaard, 'Name standardization in record linkage'.

a requirement since it is relatively well know that the spelling of names can vary over time and between sources, in part as a result of enumeration practises (a name would often be spoken to the person responsible for enumerating the source document and this spelling could be interpreted in a number of different forms) and in part due to source transcription errors.

In order to overcome this problem a number of solutions have been adopted. The most commonly used form of name standardisation in historical linking projects has been Soundex, either in its original form or in a modified version. In effect this is a phonetic coding system in which vowels are suppressed and consonants are coded into a numeric form. However, much of the literature on historical record linkage has focused on a discussion of the adequacies of Soundex and how it needs to be modified or alternative approaches used for historical records. Other letter coding schemes that have been used include the New York State Identification Intelligence System (NYSIIS).

Interestingly, in linguistics the problem of variant spellings of similar words has often used a rather different approach, in particular, so-called *n*-gram matching.[60] This is essentially based on the principle of pattern matching, testing to see if the combination of letters in one character string (such as a name) is similar to that in another. Such techniques, perhaps surprisingly, have not been utilised much in historical research. One method that has drawn on the principal of pattern matching is the so-called Guth algorithm, designed in part to take account of 'ethnic' (non-Anglo-Saxon) names which do not always perform well to standard phonetic coding, such as Soundex.

One of the problems with both phonetic coding systems such as Soundex and NYSIIS, and the pattern-matching approach of Guth is that they in effect produce binary outcomes. In other words, the substituted phonetic code in one file will either link to exactly the same substituted phonetic code in another file, or it will not. Equally, the way in which the Guth algorithm is constructed means that the comparison between strings (names) in different files is either deemed to be 'true' or 'false'. In both cases, there are no 'maybe's'.

To overcome this problem, a number of pattern matching based utilised in linguistic research and probability based statistical medical record linkage attempt to produce a 'likeness' or probability score, usually on a scale of 0 to 1 with 1 being a perfect match (all letters being present in both character strings being compared, and in the same order) and 0 being a complete non-match (both strings being composed of entirely different characters). One historical project that has addressed this approach is the PRDH (see section 6.2.4) based on seventeenth and eighteenth-century French Canadian parish registers. This has developed a 'similarity' index between character strings. Within computing linguistics a similar measure that has been devised is the so-called Levenshtein Method which in comparing two strings, calculates an index of similarity based

[60] See Zobel and Dart, 'Finding approximate matches in large lexicons' and Rogers and Willett, 'Searching of historical word forms in text databases'.

on the number of character insertions or deletions that would be required to convert one string into the other.

Although potentially powerful, a problem of applying such similarity algorithms, as with the Guth algorithm, is that in order to compute them, each name in one source has to be compared with each name in the other sources to be linked. This comes with an overhead in computing terms and may explain why such approaches have rarely been used in historical projects. Thus, to take a hypothetical case in which an attempt was being made to link 5,000 individuals listed in one census to a list of 5,000 in a second census. Where phonetic coding schemes are applied (such as Soundex) all that is required is to pass through each dataset once and apply the standardised code before linking – a total of 10,000 computations. However, in order to produce a similarity index, assuming that each list contained, say 2,000 unique surname variations, this would require 4 million computations (2,000*2,000) before linkage could be attempted. This may be appropriate for small datasets, but given that in the 1881 census alone there are some 400,000 recorded surname variants, this does raise non-trivial, yet not insurmountable, issues of computing capacity.

Having reviewed the literature across different disciplines on the problem of name standardisation, the overwhelming conclusion was that, although claims are made about the superiority of method *x* over method *y*, there simply is no optimum solution. All approaches seem to have a number of advantages and disadvantages. Thus the main recommendation for the VPS regarding name standardisation is not to use any one single method (as is the case in nearly all historical record linkage projects) but to use multiple methods and to base linkage on the combination of outcomes.

As a by-product of reviewing the problem of name standardisation, a new method which takes as it starting point the Guth algorithm, but develops this further by applying similarity index approaches is proposed. The result is a new approach which is a hybrid between the pattern matching and similarity index approaches. The key advantage to this new approach is that the outcome is a relative, rather than binary, score, in which the higher the number, the better the likeness between the two strings being compared. This is outlined below.

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| **Surname: Robbins** | | | | | | | |
| Singles | R, 1 | O, 2 | B, 3 | B, 4 | I, 5 | N, 6 | S, 7 |
| Pairs | RO, 1 | OB, 2 | BB, 3 | BI, 4 | IN, 5 | NS, 6 | |
| Triplets | ROB, 1 | OBB, 2 | BBI, 3 | BIN, 4 | INS, 5 | | |
| **Surname: Robyns** | | | | | | | |
| Singles | R, 1 | O, 2 | B, 3 | Y, 4 | N, 5 | S, 6 | |
| Pairs | RO, 1 | OB, 2 | BY, 3 | YN, 4 | NS, 5 | | |
| Triplets | ROB, 1 | OBY, 2 | BYN, 3 | YNS, 4 | | | |

The new method starts by breaking down each character string being compared into single letters, pairs of letters and triplets, with the starting position of each being recorded. Thus the surnames ROBBINS and ROBYNS would be broken down as shown above.

Then each of these is compared in turn, first to find if the single letters, pairs and triplets identified in the first name (ROBBINS) occur in the second (ROBYNS), and if so, the distance between their positions within the strings. Distances are calculated by taking the difference of the starting positions, dividing this by the length of the longer string, and taking the result away from 1. Thus, in the example above, the R in ROBBINS would obtain a distance score of 1.00, since this matches with an R in ROBYNS in exactly the same position, while both the N and S in ROBBINS would obtain a distance score of 0.86, since these are both found in ROBYNS but are 1 position removed (=1 – (1/7)). Comparing the two example names therefore produces the following distance scores:

**Comparing Robbins with Robyns**

| | Distance scores | | | | | | | Total score | Maximum total score |
|---|---|---|---|---|---|---|---|---|---|
| Singles | 1.00 | 1.00 | 1.00 | 0.86 | 0.00 | 0.86 | 0.86 | 5.58 | 7.00 |
| Pairs | 1.00 | 1.00 | 0.00 | 0.00 | 0.00 | 0.86 | | 2.86 | 6.00 |
| Triplets | 1.00 | 0.00 | 0.00 | 0.00 | 0.00 | | | 1.00 | 5.00 |
| Totals | | | | | | | | 9.44 | 18.00 |

Likeness score = (total score / maximum total score) * 100
= (9.44 / 18) * 100
= 52.44

The aggregated scores for single letters, pairs and triplets are then summed together and divided by the maximum total score to produce the likeness index.

In addition, to this new likeness score, a series of programmes have been developed to automatically apply standard Soundex, Guth and NYSIIS codes to forenames and surnames. Also produced is an alternative Soundex code in which the first letter of a name is treated in the same way as subsequent letters in order to try and overcome the over reliance in standard Soundex on the accuracy of the initial letter in a name.[61] Lastly, programmes have also been written

[61] In standard Soundex, the first letter of a name is used directly in the resulting Soundex code. This is potentially a problem with nineteenth-century souces and transcriptions since the initial capital letter can be subject to misinterpretation. Thus Hall might be misread as Wall.

to produce the three similarity indexes proposed by Snae and Diaz,[62] since these appear to combine well and compliment the new likeness score described earlier.

The outcome of this exercise can be seen in Table 5.1, which produces the various codes and indexes for a selected number of surnames and forenames. This reinforces the point made earlier than there is no one name standardisation system that is better than the others. All have strengths and weaknesses. Thus a key recommendation of this project is that the VPS should *embed* multiple name standardisations in the linking process and attempt to establish pairings and links based on the combination of outcomes.

There is one further problem with the standardisation of name, specific to forenames that needs to be addressed. This is the fact that forenames can be corrupted in a form which no standardisation method can adequately resolve. Thus, an Elizabeth could also be called Libby, Janet could be called Jessie, John could be called Jack, and so on. Realistically, the only practical way to resolve this kind of naming problem is to produce a look-up table with all the potential matching pairs of names, in effect a thesaurus of names. This should be produced as part of a pre-linking exercise for reference by the linking algorithms.[63]

Notes to Table 5.1 and 5.2 (shown below):

| | |
|---|---|
| Surname1/Forename1 | Surname or Forename string – original |
| Surname2/Forename2 | Surname or Forename string – comparison |
| VPS Score | New pattern matching score |
| Guth | Guth algorithm outcome |
| LIG1 | ISG/Guth composite index of similarity |
| LIG2 | Levenshtein index of similarity |
| LIG3 | Levenshtein/ISG/Guth composite index of similarity |
| Soundex1 | Standard Soundex for first string |
| Soundex1b | VPS modified Soundex for first string |
| Soundex2 | Standard Soundex for second string |
| Soundex2b | VPS modified Soundex for second string |
| NYSIIS1 | New York State Identification Intelligence Systems for first string (De Bueno, 1966) |
| NYSIIS2 | New York State Identification Intelligence Systems for second string |

[62] Snae and Diaz, 'An interface for mining genealogical nominal data'.

[63] Indeed, as part of the project, a skeleton thesaurus for surnames and forenames has been produced based on the 1851 Anderson sample and the 1881 census. This would need to be added to for the other census years which form part of the VPS.

Tab. 5.1: Example standardisation of surnames

| | Surname1 | Surname2 | VPS Score | Guth | LIG1 | LIG2 | LIG3 |
|---|---|---|---|---|---|---|---|
| 1 | AINSCOMBE | ANSCOMB | 0.60 | TRUE | 0.78 | 0.78 | 0.88 |
| 2 | AINSCOMBE | ANSCOMBE | 0.71 | TRUE | 0.89 | 0.89 | 0.94 |
| 3 | AINSCOMBE | BRANSCOMB | 0.56 | FLASE | 0.78 | 0.64 | 0.88 |
| 4 | AINSCOMBE | DASCOMBE | 0.59 | TRUE | 0.78 | 0.70 | 0.88 |
| 5 | AINSCOMBE | DUNSCOMBE | 0.75 | TRUE | 0.78 | 0.64 | 0.88 |
| 6 | AINSCOMBE | HANSCOMB | 0.66 | TRUE | 0.78 | 0.70 | 0.88 |
| 7 | AINSCOMBE | LIPSCOMBE | 0.67 | TRUE | 0.78 | 0.64 | 0.88 |
| 8 | AINSCOMBE | LUSCOMBE | 0.56 | FLASE | 0.67 | 0.55 | 0.80 |

| | Soundex1 | Soundex1b | Soundex2 | Soundex2b | NYSIIS1 | NYSIIS2 |
|---|---|---|---|---|---|---|
| 1 | A525 | 5251 | A525 | 5251 | ANSCANB | ANSCANB |
| 2 | A525 | 5251 | A525 | 5251 | ANSCANB | ANSCANB |
| 3 | A525 | 5251 | B652 | 1652 | ANSCANB | BRANSCANB |
| 4 | A525 | 5251 | D251 | 3251 | ANSCANB | DASCANB |
| 5 | A525 | 5251 | D525 | 3525 | ANSCANB | DANSCANB |
| 6 | A525 | 5251 | H525 | 5251 | ANSCANB | HANSCANB |
| 7 | A525 | 5251 | L125 | 4125 | ANSCANB | LAPSCANB |
| 8 | A525 | 5251 | L251 | 4251 | ANSCANB | LASCANB |

Tab. 5.2: Example standardisation of forenames

| | Forename1 | Forename2 | VPS Score | Guth | LIG1 | LIG2 | LIG3 |
|---|---|---|---|---|---|---|---|
| 1 | JULIA | JULIAN | 0.80 | TRUE | 0.83 | 0.83 | 0.91 |
| 2 | JULIA | JULIANA | 0.71 | TRUE | 0.71 | 0.71 | 0.83 |
| 3 | JULIA | JULIANNA | 0.60 | TRUE | 0.63 | 0.63 | 0.77 |
| 4 | JULIA | JULIE | 0.75 | FLASE | 0.80 | 0.67 | 0.89 |
| 5 | JULIA | JULIET | 0.60 | FLASE | 0.67 | 0.57 | 0.80 |
| 6 | JULIA | JULIUS | 0.63 | FLASE | 0.67 | 0.57 | 0.80 |

| | Soundex1 | Soundex1b | Soundex2 | Soundex2b | NYSIIS1 | NYSIIS2 |
|---|---|---|---|---|---|---|
| 1 | J400 | 2400 | J450 | 2450 | JAL | JALAN |
| 2 | J400 | 2400 | J450 | 2450 | JAL | JALAN |
| 3 | J400 | 2400 | J450 | 2450 | JAL | JALAN |
| 4 | J400 | 2400 | J400 | 2400 | JAL | JAL |
| 5 | J400 | 2400 | J430 | 2430 | JAL | JALAT |
| 6 | J400 | 2400 | J420 | 2420 | JAL | JAL |

### *5.2.2. Other required standardisations*

Although the existing literature concentrates overwhelmingly on the standardisation of surnames, in order to perform the required record linkage, a VPS would also need to standardise a number of other information fields: namely, occupations, birthplace, place of residence, relationship to head of household, and marital condition. In fact, regardless of linkage, this task would need to be undertaken for analytical purposes as well, since in the raw nineteenth-century census data these are all recorded as un-standardised character strings. Due to the number of unique variances recoded in the 'raw' census data this is a relatively straight-forward for marital condition and relationship to household head, but rather more complex and time consuming for occupations and especially place of birth.[64] For example in the case of the 1881 census the number of unique strings occurring in the data are as follows:

- relationship to household head – 17,167
- occupation – 1,606,585
- county of birth – 7,116
- parish of birth – 855,374.

This task has already been carried out in earlier projects at the University of Essex for the Anderson 1851 sample and 1881, and is currently underway for the 1861 census of England and Wales. Having standardised the information for one census makes the task of doing it for a second very much easier and quicker,[65] but the importance of this pre-linking data processing task and the resources needed should not be under estimated.[66] Indeed, it could be strongly argued that the quality of any linkage on nineteenth-century sources is proportional to the effort made in pre-processing (cleaning and standardising) the underlying input data. Thus, in the case of the VPS, it is important to realise that the task of preparing the data could take as long, if not longer that the actual process of linking *per se*.

---

[64] Although the number of unique strings for parish of birth tends to be less than for occupation, standardising parish of birth represents greater challenges as the information is more ambiguous, in that it can contain non-parish names, such as hamlets, and in that counties can contain more than one parish with the same name.

[65] For example, as a direct result of the current project the 1861 census data for England and Wales have been made available by findyourpast.com. Using the existing look-up tables and coding dictionaries created for the early 1881 project, it proved possible to code some occupations for 92% of the population on a single pass, relationships for 99%, marital condition for 99%, county of birth for 91% and parish of birth for 83%. Although these figures are high when considered in relation to the proportion of the total population coded, it still leaves a large number of un-coded strings to be standardised. Thus, to standardise occupations for the remaining 8% of the population required the coding of some 500,000 extra occupational titles.

[66] This often neglected point is also made in Schofield, 'Automatic family reconstitution'.

In standardising the non-name information in the underlying census data it is also important both for analytical and linking reasons to provide, where possible, multiple standardisations. Thus, in the case of occupations it would be important to produce activity level codes as well as sectional level codes (see section 6.4). In the case of birthplaces, although previous codings of the 1851 Anderson sample and the 1881 census have concentrated on standardising parish and county titles, for the VPS it is recommended that this task is taken a stage further and that Ordnance Survey grid references are attached to the data. This would not only facilitate the mapping of the underlying data, but would also enable relative rather than absolute comparisons to be made in the linking process and subsequent analysis of the data. This is not an impossible task, but one that would have important resource implications.

## 5.3. Multiple links

One of the concerns that is raised in discussions regarding historical record linkage is the problem of multiple links. In particular, that it might prove impossible to resolve due to the large number of records having common (high frequency) surnames and forenames. In other words, there are far too many John Smith's, many of whom also name their son(s) John Smith, and it is difficult, if not impossible, to distinguish between them. Despite such claims, literature on the frequency distribution of either forenames or surnames is minimal.[67] In consequence, the following section examines existing census data for 1851 and 1881 in an attempt to measure the extent of this potential problem.

### *5.3.1. Forename and surname frequencies*

Using the Anderson sample for 1851 and the computerised version of the entire l881 census, frequency counts of surnames, male forenames and female forenames were made for both years (having first cleaned the data). The twenty most popular surnames in both years are given in Table 5.3 together with the proportion of the total population that they account for. Interestingly, there is a relatively large degree of correspondence between the two years, with sixteen surnames appearing in the top twenty in both years. The top three names – Smith, Jones and Williams – are exactly the same in both years, accounting for very similar proportions of the total population. Equally, collectively, the top twenty surnames accounted for 10.98 and 10.45 per cent of the overall population in 1851 and 1881 respectively. The fact that some 10 per cent of the population shared just twenty surnames may, indeed, give cause for concern, however, concentrating on just the most popular names can be misleading.

[67] For exceptions see, Redmonds, *Christian names in local and family history* and Rogers, *The surname detective*.

Certainly a small number of surnames were (and still are) very popular, but the other dominant feature of the surname frequency distribution is the sheer number of different surnames recorded in the census data. Turning the issue round the other way, in 1851 and 1881 90 per cent of the population shared a combined total of 34,607 and 445,706 surnames, respectively.[68] The key point is that while there are a small number of high frequency surnames, the overwhelming majority of the population shared medium and low frequency surnames, which pose fewer concerns for record linkage. Overall the number of persons per surname in the 1851 2 per cent sample is just 18 and still only around 104 in the case of the entire 1881 census. When placed in this broader context the issue of popular surnames seems less of a problem.

Tab. 5.3: Twenty most popular surnames in 1851 and 1881

| | 1851 (Anderson sample) | | | |
|---|---|---|---|---|
| Rank | Surname | n. | % | Cumulative % |
| 1 | SMITH | 5,557 | 1.41 | 1.41 |
| 2 | JONES | 4,782 | 1.21 | 2.82 |
| 3 | WILLIAMS | 3,097 | 0.79 | 4.03 |
| 4 | BROWN | 2,658 | 0.67 | 4.82 |
| 5 | TAYLOR | 2,337 | 0.59 | 5.49 |
| 6 | THOMAS | 1,901 | 0.48 | 6.08 |
| 7 | DAVIES | 1,807 | 0.46 | 6.56 |
| 8 | WILSON | 1,734 | 0.44 | 7.02 |
| 9 | ROBERTS | 1,581 | 0.40 | 7.46 |
| 10 | EVANS | 1,542 | 0.39 | 7.86 |
| 11 | WOOD | 1,318 | 0.33 | 8.25 |
| 12 | WALKER | 1,260 | 0.32 | 8.59 |
| 13 | LEWIS | 1,231 | 0.31 | 8.91 |
| 14 | CLARK | 1,184 | 0.30 | 9.22 |
| 15 | WRIGHT | 1,166 | 0.30 | 9.52 |
| 16 | ROBINSON | 1,157 | 0.29 | 9.82 |
| 17 | JACKSON | 1,153 | 0.29 | 10.11 |
| 18 | THOMPSON | 1,146 | 0.29 | 10.40 |
| 19 | JOHNSON | 1,134 | 0.29 | 10.69 |
| 20 | HARRIS | 1,124 | 0.28 | 10.98 |

[68] The figure for 1851 takes into account only the 2 per cent sample, while the figure for 1881 includes 100 per cent of the population. The 1881 figure, however, does not represent 'real' surnames, but rather surname 'strings' which include a large number of typing errors and miss-keying. The total number of 'real' surnames is probably nearer 250,000.

Cont. Tab. 5.3:

| | 1881 | | | |
|---|---|---|---|---|
| Rank | Surname | n. | % | Cumulative % |
| 1 | SMITH | 370,104 | 1.42 | 1.42 |
| 2 | JONES | 336,690 | 1.29 | 2.71 |
| 3 | WILLIAMS | 213,337 | 0.82 | 3.52 |
| 4 | TAYLOR | 172,507 | 0.66 | 4.18 |
| 5 | BROWN | 156,442 | 0.60 | 4.78 |
| 6 | DAVIES | 151,476 | 0.58 | 5.36 |
| 7 | EVANS | 129,859 | 0.50 | 5.86 |
| 8 | THOMAS | 122,595 | 0.47 | 6.33 |
| 9 | ROBERTS | 110,763 | 0.42 | 6.75 |
| 10 | WILSON | 99,994 | 0.38 | 7.13 |
| 11 | JOHNSON | 98,780 | 0.38 | 7.51 |
| 12 | ROBINSON | 94,065 | 0.36 | 7.87 |
| 13 | WRIGHT | 87,964 | 0.34 | 8.21 |
| 14 | WOOD | 87,198 | 0.33 | 8.54 |
| 15 | WHITE | 86,624 | 0.33 | 8.87 |
| 16 | THOMPSON | 85,150 | 0.33 | 9.20 |
| 17 | HALL | 83,620 | 0.32 | 9.52 |
| 18 | WALKER | 82,841 | 0.32 | 9.84 |
| 19 | GREEN | 81,694 | 0.31 | 10.15 |
| 20 | EDWARDS | 81,559 | 0.31 | 10.46 |

Turning to forenames, the top twenty female and male names in both years is provided in Tables 5.4 and 5.5. Unsurprisingly, due to the changing fashion of given christen names, there is greater change in the top twenty forenames over time than is evident with surnames, although the top five forenames for both males and females are the same in both years. From a linkage point of view, what is more worrying is the fact that in the nineteenth century children were named from a quite small stock of forenames. Although the size of the forename pool does increase over time[69] a striking feature is the fact that in 1851 73.95 per cent of females and 81.93 per cent of males shared just twenty forenames, the respective figures for 1881 declining to 68.8 and 78.72 per cent. In both years around one in three females where called Mary, Elizabeth or Ann(e), and one in three males, John, William or Thomas. Thus, the relatively

[69] A trend that continues to the present day.

small stock of given forenames, could potentially pose a greater linking problem for the VPS than surnames.

Tab. 5.4: Twenty most popular female forenames in 1851 and 1881

| | 1851 | | | |
|---|---|---|---|---|
| Rank | Forename | n. | % | Cumulative% |
| 1 | MARY | 33,147 | 16.38 | 16.38 |
| 2 | ELIZABETH | 25,295 | 12.50 | 28.89 |
| 3 | ANN | 18,410 | 9.10 | 37.99 |
| 4 | SARAH | 15,277 | 7.55 | 45.54 |
| 5 | JANE | 11,368 | 5.62 | 51.16 |
| 6 | MARGARET | 6,911 | 3.42 | 54.57 |
| 7 | HANNAH | 5,376 | 2.66 | 57.23 |
| 8 | EMMA | 3,933 | 1.94 | 59.17 |
| 9 | ELLEN | 3,829 | 1.89 | 61.06 |
| 10 | MARTHA | 3,525 | 1.74 | 62.81 |
| 11 | CATHERINE | 3,087 | 1.53 | 64.33 |
| 12 | MARIA | 2,882 | 1.42 | 65.76 |
| 13 | ISABELLA | 2,443 | 1.21 | 66.97 |
| 14 | CHARLOTTE | 2,336 | 1.15 | 68.12 |
| 15 | JANET | 2,203 | 1.09 | 69.21 |
| 16 | HARRIET | 2,152 | 1.06 | 70.27 |
| 17 | CAROLINE | 1,970 | 0.97 | 71.25 |
| 18 | SUSAN | 1,902 | 0.94 | 72.19 |
| 19 | ALICE | 1,805 | 0.89 | 73.08 |
| 20 | AGNES | 1,760 | 0.87 | 73.95 |

| | 1881 | | | |
|---|---|---|---|---|
| Rank | Forename | n. | % | Cumulative% |
| 1 | MARY | 1,826,814 | 13.65 | 13.65 |
| 2 | ELIZABETH | 1,429,003 | 10.67 | 24.32 |
| 3 | ANN | 1,012,558 | 7.56 | 31.88 |
| 4 | SARAH | 909,858 | 6.80 | 38.68 |
| 5 | JANE | 533,795 | 3.99 | 42.67 |
| 6 | ELLEN | 404,529 | 3.02 | 45.69 |
| 7 | ALICE | 380,614 | 2.84 | 48.53 |
| 8 | EMMA | 373,012 | 2.79 | 51.32 |
| 9 | MARGARET | 325,859 | 2.43 | 53.75 |
| 10 | EMILY | 289,189 | 2.16 | 55.91 |
| 11 | HANNAH | 279,909 | 2.09 | 58.00 |
| 12 | MARTHA | 228,246 | 1.70 | 59.71 |
| 13 | LOUISA | 168,896 | 1.26 | 60.97 |

Cont. Tab. 5.4:

| | | | | |
|---|---|---|---|---|
| 14 | HARRIET | 164,944 | 1.23 | 62.20 |
| 15 | CATHERINE | 163,184 | 1.22 | 63.42 |
| 16 | MARIA | 152,868 | 1.14 | 64.56 |
| 17 | EDITH | 144,962 | 1.08 | 65.64 |
| 18 | CHARLOTTE | 143,665 | 1.07 | 66.72 |
| 19 | FLORENCE | 140,902 | 1.05 | 67.77 |
| 20 | FANNY | 138,302 | 1.03 | 68.80 |

Tab. 5.5: Twenty most popular male forenames in 1851 and 1881

| | 1851 | | | |
|---|---|---|---|---|
| Rank | Forename | n. | % | Cumulative% |
| 1 | JOHN | 30,999 | 16.09 | 16.09 |
| 2 | WILLIAM | 28,934 | 15.02 | 31.12 |
| 3 | THOMAS | 17,218 | 8.94 | 40.05 |
| 4 | JAMES | 17,161 | 8.91 | 48.96 |
| 5 | GEORGE | 11,756 | 6.10 | 55.07 |
| 6 | HENRY | 7,197 | 3.74 | 58.80 |
| 7 | JOSEPH | 6,935 | 3.60 | 62.40 |
| 8 | ROBERT | 6,927 | 3.60 | 66.00 |
| 9 | CHARLES | 5,644 | 2.93 | 68.93 |
| 10 | RICHARD | 4,411 | 2.29 | 71.22 |
| 11 | EDWARD | 4,055 | 2.11 | 73.33 |
| 12 | SAMUEL | 3,581 | 1.86 | 75.19 |
| 13 | DAVID | 3,391 | 1.76 | 76.95 |
| 14 | ALEXANDER | 1,858 | 0.96 | 77.91 |
| 15 | ALFRED | 1,540 | 0.80 | 78.71 |
| 16 | PETER | 1,459 | 0.76 | 79.47 |
| 17 | BENJAMIN | 1,221 | 0.63 | 80.10 |
| 18 | DANIEL | 1,218 | 0.63 | 80.73 |
| 19 | FREDERICK | 1,180 | 0.61 | 81.35 |
| 20 | FRANCIS | 1,125 | 0.58 | 81.93 |

| | 1881 | | | |
|---|---|---|---|---|
| Rank | Forename | n. | % | Cumulative% |
| 1 | WILLIAM | 1,660,103 | 13.03 | 13.03 |
| 2 | JOHN | 1,591,008 | 12.49 | 25.53 |
| 3 | THOMAS | 955,900 | 7.51 | 33.03 |
| 4 | GEORGE | 853,713 | 6.70 | 39.74 |
| 5 | JAMES | 822,067 | 6.45 | 46.19 |
| 6 | HENRY | 658,492 | 5.17 | 51.36 |

Cont. Tab. 5.5:

| | | | | |
|---|---|---|---|---|
| 7 | CHARLES | 490,280 | 3.85 | 55.21 |
| 8 | JOSEPH | 439,592 | 3.45 | 58.66 |
| 9 | ROBERT | 332,016 | 2.61 | 61.27 |
| 10 | EDWARD | 297,218 | 2.33 | 63.60 |
| 11 | FREDERICK | 274,757 | 2.16 | 65.76 |
| 12 | ALFRED | 258,082 | 2.03 | 67.79 |
| 13 | ARTHUR | 239,985 | 1.88 | 69.67 |
| 14 | RICHARD | 234,206 | 1.84 | 71.51 |
| 15 | SAMUEL | 219,258 | 1.72 | 73.23 |
| 16 | WALTER | 181,266 | 1.42 | 74.65 |
| 17 | ALBERT | 165,509 | 1.30 | 75.95 |
| 18 | DAVID | 146,907 | 1.15 | 77.11 |
| 19 | FRANK | 106,168 | 0.83 | 77.94 |
| 20 | HERBERT | 99,397 | 0.78 | 78.72 |

### *5.3.2. Forename-Surname pairs*

However, for linking purposes what matters most are pairs of names – the combination of forename and surname that individuals had. Examining the census data for 1881 the ten most popular forename-surname pairs is given in Table 5.6. This shows that although nationally the number of Mary Jones, John Jones' Mary Smith's and William Smith's can be counted in tens of thousands, in overall proportionate terms they were relatively insignificant. Moreover, Table 5.7 shows that whilst a small proportion of the population, had common pairings, with many individuals having the same forename and surname, most did not. Indeed, of the males recorded in the 1851 sample, a somewhat surprising 35 per cent had a totally unique combination of forename and surname and a further 26 per cent had only two, three or four persons sharing a common forename-surname paring. Thus, whilst multiple linkages might prove to be a problem for one or two per cent of the population, more importantly, they should not for the vast majority of the rest.

Taking this one step further, in the case of John Jones, the most popular male forename-surname pairing, there are 425 John Jones' recorded in the 1851 sample. Breaking these down by age and place of birth there are only ten individuals with exactly the same name, age and place of birth – those aged between 0 and 1 born in Abergavenny. When these are then assigned attribute variables – parents names and birth year – all are unique. Thus, whilst at the individual level several people may share the same surname and forename, once additional information is added, especially at the household level, com-

monality becomes increasingly rare.[70] This example again emphasises the need in the case of the VPS to place linkage in the context of the household in which individuals are resident.

Tab. 5.6: Ten most popular forename-surname pairs, 1881

| Female pairs | | | | |
|---|---|---|---|---|
| | Surname | Forename | Count | % |
| 1 | JONES | MARY | 26,854 | 0.20 |
| 2 | SMITH | MARY | 24,851 | 0.19 |
| 3 | WILLIAMS | MARY | 16,877 | 0.13 |
| 4 | JONES | ELIZABETH | 16,336 | 0.12 |
| 5 | SMITH | ELIZABETH | 14,242 | 0.11 |
| 6 | DAVIES | MARY | 13,723 | 0.10 |
| 7 | SMITH | SARAH | 13,532 | 0.10 |
| 8 | TAYLOR | MARY | 12,275 | 0.09 |
| 9 | JONES | SARAH | 11,347 | 0.08 |
| 10 | EVANS | MARY | 10,892 | 0.08 |

| Male pairs | | | | |
|---|---|---|---|---|
| | Surname | Forename | Count | % |
| 1 | JONES | JOHN | 26,077 | 0.21 |
| 2 | SMITH | WILLIAM | 22,713 | 0.20 |
| 3 | SMITH | JOHN | 21,665 | 0.13 |
| 4 | JONES | WILLIAM | 21,469 | 0.13 |
| 5 | WILLIAMS | JOHN | 15,928 | 0.11 |
| 6 | JONES | THOMAS | 15,227 | 0.11 |
| 7 | WILLIAMS | WILLIAM | 13,662 | 0.11 |
| 8 | SMITH | GEORGE | 13,548 | 0.10 |
| 9 | SMITH | THOMAS | 12,634 | 0.09 |
| 10 | DAVIES | JOHN | 12,008 | 0.09 |

[70] This is in part confirmed by the work of Long who has attempted to link the Anderson 1851 sample with the complete 1881 census. Taking only males and using only name, age and birthplace information, and discarding all multiple links, he reports a linkage rate of 15 per cent. Given the proportions who will have died or emigrated during the thirty year interval between the two censuses this is an encouraging figure.

Tab. 5.7: Numbers in each male forename-surname pair, 1851 and 1881

| 1881 | | | 1851 | | |
|---|---|---|---|---|---|
| *n. within each pair* | *n. of individuals* | *% of total population* | *n. within each pair* | *n. of individuals* | *% of total population* |
| 10,000+ | 231,231 | 1.8 | 400+ | 843 | 0.5 |
| 7,500-9,999 | 120,394 | 1.0 | 300-399 | 702 | 0.4 |
| 5,000-7,499 | 232,119 | 1.8 | 200-299 | 1,583 | 0.8 |
| 2,500-4,999 | 563,496 | 4.4 | 100-199 | 3,793 | 1.8 |
| 1,000-2,499 | 1,019,654 | 8.0 | 50-99 | 10,082 | 5.3 |
| 500-999 | 1,018,882 | 8.0 | 25-49 | 12,121 | 6.3 |
| 250-499 | 1,100,577 | 8.6 | 10-24 | 22,860 | 11.9 |
| 100-249 | 1,519,050 | 11.9 | 5-9 | 22,227 | 11.6 |
| 50-99 | 1,114,586 | 8.8 | 2-4 | 49,778 | 26.0 |
| 25-49 | 1,058,658 | 8.3 | 1 | 67,938 | 35.4 |
| 10-24 | 1,258,612 | 9.9 | | | |
| 5-9 | 858,855 | 6.7 | | | |
| 2-4 | 1,201,210 | 9.4 | | | |
| 1 | 1,438,583 | 11.3 | | | |
| *Total* | 12,735,907 | 100.0 | | 191,975 | 100.0 |

### *5.3.3. Alternative solutions*

Although it is believed that the often stated problem of common surnames and forenames is actually less of a problem than sometimes perceived, in examining the potential issue of multiple links caused by common surnames and forenames further, a unique and previously unused solution is proposed for the VPS.

Most previous historical linkage exercises have been place orientated, linking data from a particular place to the same place over time. Places are chosen invariably because of the research questions to be answered, usually issues of employment, migration or demography to be addressed. Only after the place is selected and the linkage process started is it discovered that the surname-forename pools for that particular place may be unusually small. However, in the case of the VPS, as is demonstrated above, information on surname-forename pools will be known in advance, and thus it is possible to calculate the probability of each forename-surname pairing to enter into a multiple link. Thus, turning the problem around, if multiple links were to be a problem due to commonly used forenames and surnames, it would be entirely possible to re-sample the base population, *excluding* individuals from the sample with com-

mon forename-surname pairs. Such an approach would be acceptable as long as it is possible to demonstrate that those individuals/families with common forename-surname pairs are not significantly different in their employment/migratory/demographic behaviour from those with less common forename-surnames. Although a future VPS project would need to undertake further detailed work in this area, the results of a preliminary exercise based on the Anderson 1851 sample are given below.

All the 81,437 heads of household in the 1851 sample were split into one of two groups by frequency of forename-surname combination: those who shared a common forename-surname pair with twenty-five or more other heads of household nationally, and those who did not. The first group, those with common and relatively common forename-surname pairs accorded for 5.4 per cent of all heads of household. Summary demographic and household measures were then calculated on the two groups, the results of which are given in Table 5.8.

Tab. 5.8: Mean values of given characteristics or those with common and non-common forename-surname pairs: heads of household, 1851

| *Mean vale* | *Heads of household with same forename-surname as <25 other heads* | *Heads of household with same forename-surname as 25+ other heads* |
|---|---|---|
| n. of offspring | 1.81 | 2.00 |
| n. of relatives | 0.28 | 0.25 |
| n. of inmates | 0.14 | 0.14 |
| n. of servants | 0.40 | 0.32 |
| mean age | 44.39 | 44.30 |
| Total | 77,072 | 4,365 |

The results show that there may, in fact, be a significant difference between the two groups, especially in terms of the numbers of residential offspring and servants. However, it may be possible to explain this difference due to regional differences. For example, the residential patterns of the Welsh are different from the English and Scots, and Welsh households are more likely to have heads of household with common forename-surname pairs. To investigate this possibility, means were calculated by country of enumeration, and as Table 5.9, looking only at Welsh heads indicates, the differences evident in Table 5.8 are due more to regional variation that any significant different between households headed by those with common forename-surname and compared to those with less common forename-surname pairs.

Tab. 5.9: Mean values of given characteristics or those with common and non-common forename-surname pairs: Welsh heads of household, 1851

| *Mean vale* | *Heads of household with same forename-surname as <25 other heads* | *Heads of household with same forename-surname as 25+ other heads* |
|---|---|---|
| n. of offspring | 2.11 | 2.20 |
| n. of relatives | 0.26 | 0.26 |
| n. of inmates | 0.19 | 0.15 |
| n. of servants | 0.34 | 0.34 |
| mean age | 46.39 | 45.93 |
| Total | 3,461 | 1,404 |

Obviously, a future VPS project would need to explore this possibility of bias more fully, and would probably need to produce thresholds of what common forename-surname pairings to exclude from the sample on a regional (rather than national) basis, but it is suggested that such an approach would be possible without biasing the underlying sample.

An alternative, yet more sophisticated approach would be to adopt what are sometimes called 'hot-decking' methods of sampling. Thus, if the initial random sample included a household with a common excluded forename-surname pair, then a substitute household would be selected in its place using the nearest match (for example, age, household composition and occupations all similar) of a household without non-excluded forename-surname pairs.

To summarise, tests on the 1851 and 1881 census would suggest that the problem of multiple links due to common forenames and surnames may not be as great as some have suggested, especially once assigned household attribute variables are taken into consideration. Even if multiple links do prove to be a linking problem then alternative sampling methods can be applied to exclude households with individuals that would be most likely to generate unresolved multiple links. Lastly, it has to be remembered, as mentioned previously, that the main purpose of linking in a VPS would be to minimise bias rather than to maximise accuracy. Thus it is also possible to resolve multiple links by selecting within the links randomly, as long as they all have an equal probability of being the 'right' link. In the VPS, from a methodological point of view, it will make no difference if 'right' links are chosen over 'wrong' links as long as the outcome can be shown not to have biased the resulting dataset and linkage between households is internally consistent.

## 5.4. The use of indexes

For the VPS, linking on incomplete indexes of sources rather than complete transcriptions could pose a greater problem than common forenames and sur-

names. This will be a problem that affects some Scottish census data and all civil registration data for England, Wales and Scotland. The problems with using indexed data are principally threefold: first, the format of the marriage indexes; second, multiple links without attribute data; third multiple events occurring between census dates.

### *5.4.1. Marriage index formats*

The basic problem with using the civil registration indexes of marriages for the VPS is the fact that before 1912 the marriage entries in the index do not give the surname of the individual's spouse. Prior to 1852, up to four marriages were entered on each page of the register. After that (with a few exceptions) there are two marriages per page. However, each page of the register is identified by the District, Volume and Page references. Thus for the majority of the period to be covered by the VPS, any one bride could like to two possible grooms (and *vice versa*) and it is impossible from the indexes alone to tell which of the two possible grooms is the right one. This is perhaps best illustrated from the following hypothetical entries from the marriage indexes.

| Forename | Surname | District number | Volume number | Page number |
|---|---|---|---|---|
| Mary | Smith | 35 | 5 | 8 |
| Anne | Jones | 35 | 5 | 8 |
| John | Clarke | 35 | 5 | 8 |
| William | Walker | 35 | 5 | 8 |

Thus, in the example above, Mary Smith could have married either John Clarke or William Walker, as could have Anne Jones. Also note that gender is not provided on the marriage index entries.

In order to resolve this problem first gender for each person has to be assigned using forename information, then 'dummy' marriages have to created matching each possible combination. Therefore, in the example given, four dummy marriages would be generated. Each of these for a given inter-censual period would then need to be linked to women of marriageable age in the previous census so that the possible future husband's name can be used as an attribute variable, together with the district where the marriage was solemnised. Although awkward, preliminary tests suggest that this is possible.

### *5.4.2. Multiple links without attribute data*

The problem of multiple links due to common forenames and surnames has already been discussed. However, in the case of indexed sources the problem is increased due to the lack of attribute data that can be used in conjunction with

them. In the case of linking deaths and births to census data age information will be particularly important. In the case of very common forename-surname pairs, it may, however, prove necessary to inspect a number of certificates before the correct individual is identified.

### *5.4.3. Multiple events occurring between census dates*

Of the three, this is undoubtedly the greatest problem facing the VPS. This principally relates to individuals (children of sample members) who were born and subsequently died within an inter-censual period. Whilst it may be possible to link births to infant/child deaths, the problem for those even with less common surnames, in the absence of parent's names is to link these to the families in which they occurred. Again, it may prove necessary to inspect a number of certificates before the correct individual is identified.

## 5.5. A linkage strategy for the VPS

Based on this examination of the record linkage problems likely to be encountered by a future VPS project, together with a number of tests carried out with existing selected census data and the indexed civil registration data, a record linkage strategy is proposed for the VPS consisting of the following stages.

Stage 1: *Clean, format and standardise census data*
The census data received from the data suppliers will need to be cleaned, formatted and standardised. This is a major undertaking in its own right, given the volume of material envisaged. As shown in Table 5.10, the total population across all censuses in the period 1851 to 1901 is some 170 million. Of this total, some 30 million records (the 1881 census) have been processed already at the University of Essex, and a further 20 million records (the 1861 census for England and Wales) have been partially processed. The task requires standard codes to be assigned to relationship to household head, marital condition, occupation and birthplace. The enumeration geography also needs to be checked and formalised. Consistency checks then need to be run, correctly assigning genders and resolving ambiguities (e.g. 'Mother' aged five). Household boundaries need to be resolved and standardised, taking account of changing household definitions over time.[71] Once completed, relationships within the household need to be analysed to form and assign conjugal family units.

Stage 2: *Assign attribute variables*
This is essentially a continuation of stage 1 in that once household composition has been formalised, attribute variables giving details of residential kin (such as par-

[71] See, for example, the discussion on household definition in Schürer and Mills, 'Family and household structure', 281-4, Anderson, 'Standard tabulation procedures', 142-3 and Higgs, 'Structuring the past'.

ents' names, birth years, birthplaces, etc) need to be generated and written to each individual, as shown in Figure 5.1.

Stage 3: *Clean, format and standardise civil registration index data*
As with stage 1, the civil registration index data will need to be checked and formatted. As shown in Table 5.11, in total some 86,471,154 births, marriages and deaths were registered in Great Britain between 1851 and 1901. The indexed data for England and Wales and the data for Scotland from GRO(S) will need to be merged together and reformatted producing combined files for Great Britain, split by event type and census to census period, totalling fifteen databases in all, as shown in Figure 5.2. The resulting data files will the need to be checked for inconsistencies. Registration district information will also need to be standardised.

Stage 4: *Produce forename and surname thesauri*
Once the census and civil registration data have been reformatted, all forenames and surnames will need to be output from the combined data to produce two aggregated databases. These will then need to be cleaned and standardised, for example, removing redundant spaces and inverted commas from surnames such as O'Connor and making consistent the use of prefixes such as 'Mac' and 'Mc'. Thesauri will then need to be generated using multiple coding systems as described previously.

Stage 5: *Produce macro level population estimates*
This is recommended not only as a devise to measure the linkage outcomes against, but is also necessary in order to produce 'best' estimates of levels of immigration and emigration which are needed in relation to the problem of sample refreshment. It should also be noted that this exercise alone will produce an entirely new set of statistics which will in turn shed new light on the levels of nineteenth-century emigration and immigration.[72] Ideally, such statistics would be calculated at a regional as well as national level, if the resources were made available. One of the advantages of undertaking a retrospective linkage exercise at a national level is that it is possible to model the expected outcomes using aggregate population data. Much can be done in this regard using the published reports of the Registrar Generals, but combining these with the tabulated individual level census and index data will enable spreadsheets to be produced giving the expected numbers of individuals in 'observation' by single year of age, year by year, together with the related number of vital events. This has already been done in part for England and Wales as a feasibility test.

[72] Working purely with published aggregate data Baines, for example, has produced estimates of net migration, but combining the macro and 100 per cent micro data would allow this work to be taken one step further. Baines, *Migration in a mature economy*.

Stage 6: *Calculate internal probabilities*
Once the census and civil registration data have been reformatted and coded then probabilities of combinations of variables occurring within the data can be calculated. These will later guide and inform the linking process.

Stage 7: *Calculate external probabilities*
Using the new set of statistics generated under stage 5, 'changing state' probabilities should be calculated, in other words, the probability of a given individual dying, marrying, being widowed or giving birth. Likewise, probabilities of data error should be calculated and correlated with other studies, for example the probability of age mis-reporting. As with the internal probabilities these will later guide and inform the linking process.

Stage 8: *Resample the underlying 1851 base population*
Assuming that a full census database for 1851 is made available to the VPS, and assuming that the recommendation to produce an EPSEM sample is followed then a new 1851 base sample should be made, replacing the Anderson 2 per cent sample. Based on the findings of stage 6, this new sample could also eliminate sample households prone to multiple linkage problems, replacing them with non or less problematic households.

Stage 9: *Create database of dummy marriages*
Due to the problem of the marriage indexes not providing a unique marriage identifier, a database of dummy marriages will need to be created in which spouses are matched and attributed to all possible partners. In doing this the records will also need to be matched to the forename and surname thesauri in order to take account of potential variants of names.

Stage 10: *Link dummy marriages to census data*
Before 'full' linking is undertaken, the dummy marriages will need to be linked to the census data. Although a number of possibilities exist, based on tests undertaken so far, it is recommended that dummy marriages occurring between census years are matched with 'marryable' women in the previous census. The point of this exercise is to assign the name(s) of potential spouses to the census records as attribute variables, in order that possible changes of surname can be identified and taken into consideration in the main linking exercise.

Stage 11: *Match census, birth and death data to name thesauri*
This is not unlike stage 9 in that matching the data to surname and forename thesauri will create dummy entries with name alternatives that will be used in the linking exercise.

Stage 12: *Identify and link inter-censual births and deaths*
Like the linking of dummy marriages, this task is seen as a separate linking exercise. However, the reason is not to assign attribute variables, but rather to match

infant and child birth-death pairs (where both birth and death occurred between censuses) and take them out of the main databases of births and deaths. In part this is an efficiency measure, but also this new database will later need to be used to attempt to reallocate them to sample members as necessary.

Stage 13: *Develop rules for scoring*

Based on the calculated probabilities and test linking exercises, scores for closeness of match will need to be developed. In reality, it is recommended that no single set of scores is used, but that different scoring algorithms are generated and applied to the linked data, and the results compared.

Stage 14: *Link every remaining census/registration combination*

Although this sounds like a huge undertaking, once all the data preparation work has been completed, and the various probabilities and scoring systems devised, this should be a relatively straightforward computing task. However, based on the preliminary tests, given the number of records under consideration, it could well take several weeks of CPU time to complete. This is especially the case since it is recommended, as illustrated in Figure 5.3, that every census is linked to every other census. This is different from most census linking exercises which generally tend to link sequentially from one census to the next only. Although obviously easier, this approach is not recommended mainly for two reasons. First, it carries the danger of missing return migration, where a sample member could have left the country and returned, where the period of absence coincided with a census year.[73] Second, from a methodological point of view it strengthens the linkage and helps with the problem of resolving multiple links. For example, a sample member in 1851 may link to several possibilities in 1861, some of which may carry the same weighting or score. By linking 1861 to 1871 *and* 1851 to 1871 it may turn out to be the case that a weaker or equal link that would have been rejected on the basis of the 1851 to 1861 comparison alone turns out to be the strongest when viewed in combination. Thus, linking across all census years also helps to maintain internal consistency and allows for the strength of aggregate links rather than individual links.

Likewise, even if the EPSEM sample approach is rejected, and the Anderson stratified cluster sample of 1851 is used as a starting point, it is recommended that links are made across whole censuses within a single step. Again this is different from most previous census linking exercises. For example, in the case of the Kingston Local History Project, when the 1871 Kinston census was linked to the national 1881 census, linkage was carried out in a number of incremental stages. First, the Kingston returns for 1871 were linked to Kingston in 1881, then those not linked were compared to the census returns of the parent Registration District, then those still unlinked where compared to the census returns for the remainder of the county, then to the region, and finally to the rest of the country. Whilst there is some logic

[73] Baines has argued that return migration was higher than perhaps previously thought. Baines, *Migration in a mature economy*.

to this, especially in terms of computing efficiency, the danger is that it potentially biases the outcome in favour of local migration, since individuals that might have been at risk to be linked to a record from a more distant location will have been eliminated on an earlier round or pass through the data.

Stage 15: *Resolve multiple links*

Once all the linking is complete, 'pathways' through the data need to be established. Aggregate probabilities and scores will need to be calculated for each path of linked pairs within the linked data. Equally competing or multiple links will need to be identified and then deconstructed or resolved. The generally accepted method for do this is to identify the pathway or chain of links which has the highest probability/score combination.[74] This is then 'banked' and all other pairings that conflict with this are deleted or eliminated from the database. Then the next highest pathway is chosen and treated in the same way, and so on until all the links are resolved.[75] An alternative method that has been proposed is to select the combination of pathways that results in the maximum total number of pathways, even though this may lead to some 'incorrect' decisions.[76] This has been suggested for family reconstitution (parish register) data where it is often important to create the maximum number of reconstituted families as possible for analytical purposes. However, in the case of the VPS, the criteria are rather different in that the main purpose is to link a specified number of predetermined individuals and households. Thus, while it may prove useful to investigate this second approach, it will most probably offer little to the VPS. Lastly, it is worth mentioning that an alternative approach which has been proposed is simple to ignore and discard more complex multiple links. Ferrie, for example, discards all individuals with more than ten potential matches. [77] Whilst such an approach is not advocated for the VPS, this may be relevant if the suggestion for a 'complete' linkage, discussed below, were to be taken forward.

Stage 16: *Check against 'expected' outcomes*

Unlike most historical linkage exercises, especially family reconstitutions based on pre-1837 parish register information, given that the VPS will be based on a nationally representative sample, it will be possible to calculate in advance a number of expected outcomes and the probabilities associated with them, as described in stage 7. Thus the initial resolution of the linking process can be compared against these statistical expectations. Depending on the outcome of this exercise scores can then be adjusted, new pathways formed and resolved in order to find a better 'fit' with the expected result.

---

[74] In the case of separate pathways sharing identical probabilities and scores then one is selected randomly.

[75] See, for example, Schürer, 'Historical demography', Schürer *et al*, 'Theory and methodology' and Schofield, 'Automatic family reconstitution'.

[76] Bouchard, 'The processing of ambiguous links'.

[77] Ferrie, 'A new sample of males'. See also, Long, 'Urbanisation, internal migration and occupational mobility' in which all multiple links are discarded.

Stage 17: *Re-link as necessary*

This is in part a continuation of stage 16, but it will also inevitably be the case that some sample members are just 'lost'. In other words attrition will occur as a result of no satisfactory link being made. For such unlinked sample members it will be necessary to re-specify the linking model and try gain.[78] Realistically this is an attempt to compensate for and overcome data error. For example, it will probably be prudent on the first linking pass through the data to form links on a fixed number of characteristics: match on forename or forename alternative, match on surname or surname alternative, match on gender, birth year within plus or minus ten years. However, because of data error some 'true' links may not meet these criteria. Age might be mis-transcribed or mis-enumerated by more than ten years; gender may be incorrectly assigned;[79] forenames, for quite legitimate reasons, might change over time.[80] An example has even been found where an enumerator gave an entire family a completely different surname by mistake. Thus, it will be necessary to relax or re-specify the linking criteria for sample members unlinked after the first pass through the data, for example by widening the birth year threshold and ignoring names altogether, concentrating instead on other fields and attribute variables, and attempt to create new pathways.

Stage 18: *Reconcile data*

Re-specification of the model and re-linking should then continue until both an acceptable level of attrition is achieved and the net result is internally consistent and corresponds to the expected outcome within statistical limits of variation. Once complete, probability or confidence measures on each pair of links will need to be calculated and written to the data.

---

[78] A similar stage is carried out in the case of the Kingston Local History Project, for example.

[79] It is important to realise that the nineteenth-century censuses did not include a specific question on gender. Instead, enumerators were required to enter the ages of males and females in two separate columns. It is certainly the case that sometimes ages were entered in the wrong column. Whilst checks for this try of error will be made during stage 1, comparing implicit gender with forename and relationship for example, rare mistakes can still occur.

[80] For example, individuals may chosen to start using a middle name as there preferred name of reference than their first or primary Christian name. In this context, it is interesting to note that Queen Victoria was actually Christened 'Alexandrina Victoria'. As a young child the future Queen was called Drina, an abbreviation of her first name, and it was only after her ascension to the throne in 1837 that she adopted her second name as her preferred name. If it were not for this, this publication might be referring to the APS rather than the VPS! See Schürer and Dillon, 'What's in a name?'.

Tab. 5.10: Population totals in the censuses of 1851 to 1901

| | Population | | |
|---|---|---|---|
| | England and Wales | Scotland | Great Britain |
| 1851 | 17,927,609 | 2,888,742 | 20,816,351 |
| 1861 | 20,066,224 | 3,062,294 | 23,128,518 |
| 1871 | 22,712,266 | 3,360,018 | 26,072,284 |
| 1881 | 25,974,439 | 3,735,573 | 29,710,012 |
| 1891 | 29,002,525 | 4,025,647 | 33,028,172 |
| 1901 | 32,527,843 | 4,472,103 | 36,999,946 |
| Total | 148,210,906 | 21,544,377 | 169,755,283 |

Tab. 5.11: Number of civil registration events, 1851 to 1901

| | Births | | |
|---|---|---|---|
| | England and Wales | Scotland | Great Britain |
| 1851-1861 | 6,487,297 | 641,348 | 8,089,578 |
| 1861-1871 | 7,536,686 | 1,125,231 | 8,661,917 |
| 1871-1881 | 8,604,710 | 1,233,990 | 9,838,700 |
| 1881-1891 | 8,798,371 | 1,251,910 | 10,050,281 |
| 1891-1901 | 9,159,208 | 1,281,625 | 10,440,833 |
| Total | 40,586,272 | 6,495,037 | 47,081,309 |
| | Deaths | | |
| | England and Wales | Scotland | Great Britain |
| 1851-1861 | 4,226,568 | 393,279 | 4,619,847 |
| 1861-1871 | 4,811,678 | 708,910 | 5,520,588 |
| 1871-1881 | 5,178,330 | 760,201 | 5,938,531 |
| 1881-1891 | 5,268,829 | 746,314 | 6,015,143 |
| 1891-1901 | 5,565,339 | 780,902 | 6,346,241 |
| Total | 25,050,744 | 3,389,606 | 28,440,350 |
| | Marriages | | |
| | England and Wales | Scotland | Great Britain |
| 1851-1861 | 1,602,281 | 128,117 | 1,730,398 |
| 1861-1871 | 1,772,655 | 224,562 | 1,997,217 |
| 1871-1881 | 1,962,281 | 253,977 | 2,216,258 |
| 1881-1891 | 2,058,588 | 260,378 | 2,318,966 |
| 1891-1901 | 2,388,764 | 297,892 | 2,686,656 |
| Total | 9,784,569 | 1,164,926 | 10,949,495 |
| Grand total | 75,421,585 | 11,049,569 | 86,471,154 |

Note: The figures for Scotland are only for 1855 onwards.

Fig. 5.1: Assignment of attribute variables

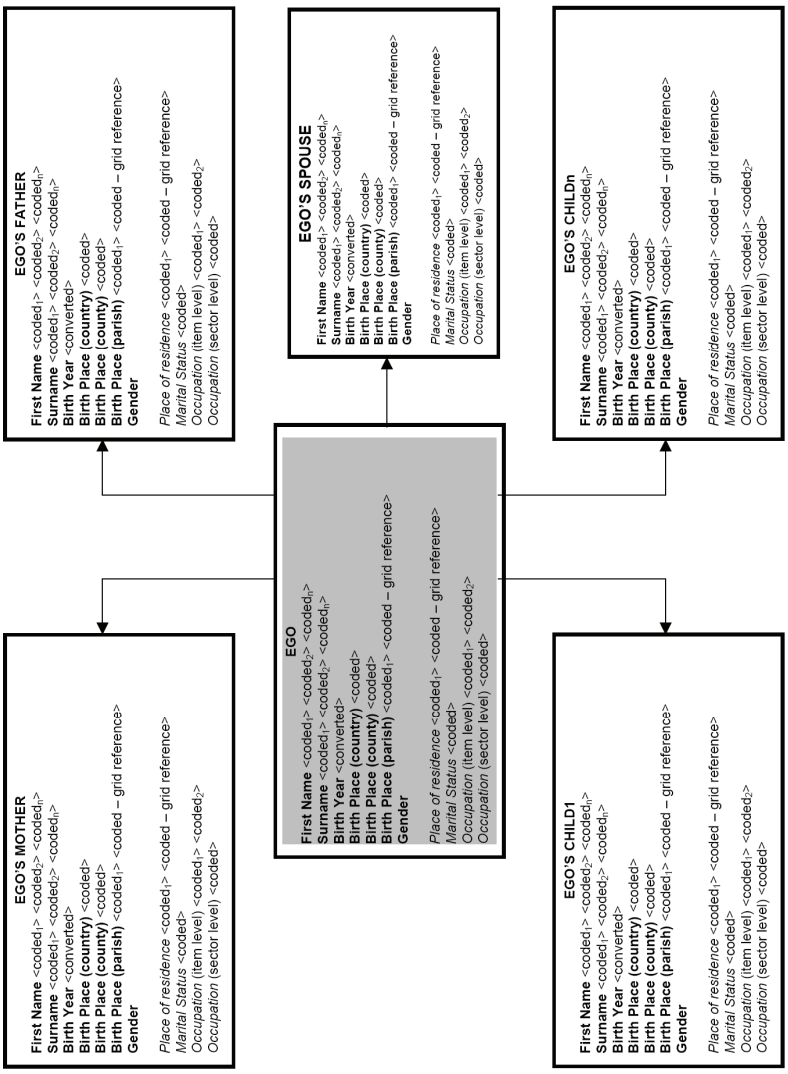


Fig. 5.2: Linkage pairs for the VPS

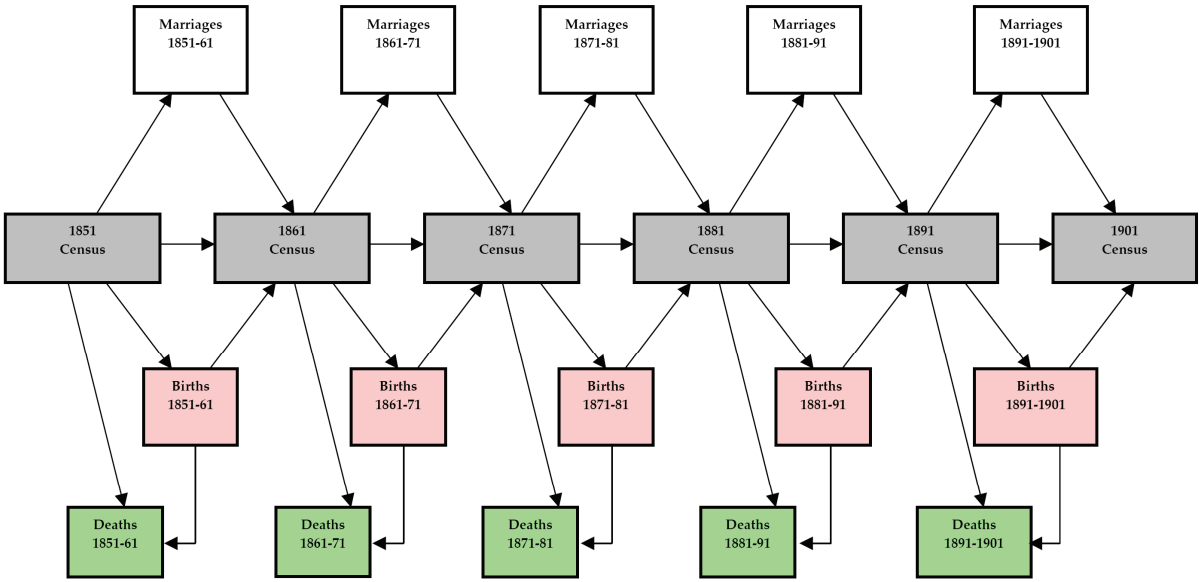
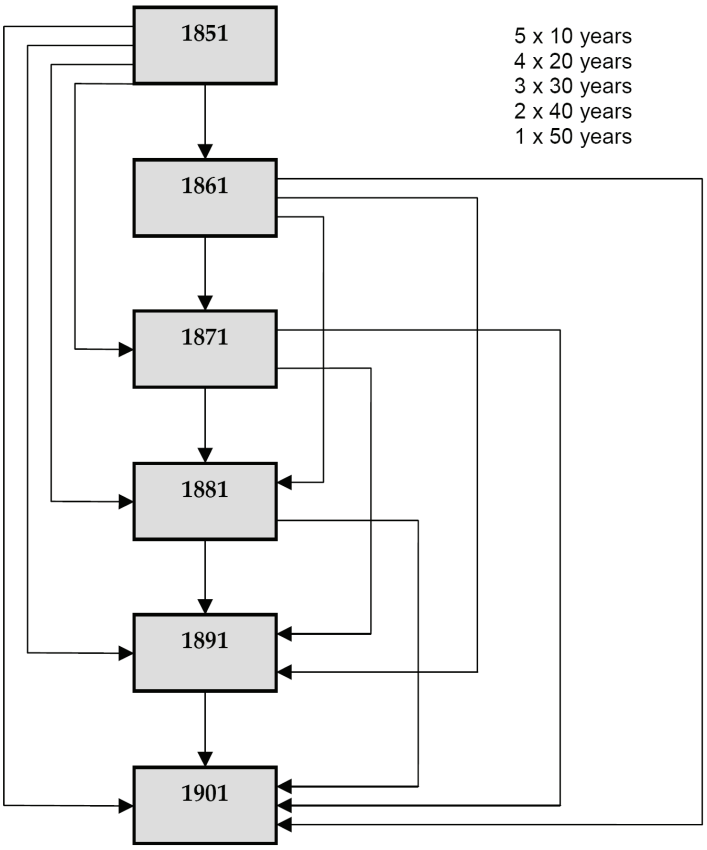


Fig. 5.3: Multiple census linkage




### *5.5.1. Other issues*

Two further issues need to be addressed. First, given the availability of full census and civil registration index data (i.e. covering 100 per cent of the population) in proportional terms it might not be a tremendous extra effort to actually attempt to link the entire population. This may sound like a fanciful suggestion, but given that the data would need to be prepared for the whole population (stage 1) the effort involved in linking the entire population, rather than a 2 per cent sample would certainly not be fifty times greater. Moreover, from a methodological point of view the idea has a number of attractions. Despite the greater numbers involved, it would actually help with the problem of

resolving multiple links in that every individual could be 'forced' to be linked, in line with defined probabilities and calculated expectations, to either a subsequent census, an intervening death or marriage or classed as an emigrant. Thus attrition would be minimised.[81] If a complete linkage were to be attempted, then the VPS could also be formulated in an entirely different way. Rather than starting with a base sample that is linked over time, a sample (or samples) could be constructed retrospectively, drawing on 'complete' reconstructions only, or those with the highest probability or confidence levels. A complete linkage exercise would also allow the possibility of users to define their own sample criteria. Samples could be area or community based, or focus on particular sub-groups in the population, for example a specific occupational group or immigrant groups, which might be too small to analyze in a 2 per cent sample. It is certainly recommended that a future VPS project should keep the options open to this possibility.

Second, it has been suggested that the VPS link backwards rather than forwards. Whilst this is a possibility, and indeed, in the case of the immigrant population and maybe in the case of marriages, is desirable in order to complete consistency checks,[82] the suggestion was made mainly from the point of view of formulating the VPS sample. The argument being that others, such as family historians and genealogists work from the present back to the past in tracing ancestors, why doesn't the VPS? It is an interesting suggestion, but upon reflection it is recommended not to take such an approach since it would led to bias and an unrepresentative sample that could not easily be corrected via refreshment. By definition, drawing a sample from the 1901 census and working backwards means that the resulting sample would be a survey of survivors, excluding all families that die out due to the failure to reproduce themselves. Thus it would not represent a true picture of either marriage, mortality or fertility.

## 5.6. Scotland

A further question that needs to be addressed is if the VPS team should create a longitudinal database for Scotland as a separate project – separate that is from England and Wales. In other words, should there in effect be not one but two Victorian Panel Surveys? In considering this question it could be argued that methodologically the two cannot be separated since Scotland's emigrants were England's immigrants, and thus the two have to be considered as a logical

[81] Attrition could never be entirely eliminated due to some under recording in the source materials. Equally, a small proportion of census enumeration books are known to 'missing', either lost or destroyed.

[82] Backwards linking is also being undertaken in the project studying Skye undertaken by Garrett and Davies at the Cambridge Group, where those identified as being born in Skye in the 1881 census are then traced back.

whole. Consequently, if a future VPS project is situated in England or Wales, which itself is open to question, while it may be politically desirable to have a separate team processing Scottish data located in Scotland working closely with GRO(S), record linkage work would need to be undertaken centrally. It should also be recognised that should a separate Scottish VPS team be established, then this would add to the costs of the project, since, for example, additional layers of project co-ordination and management would need to be put in place. Regardless of this it is clear that any future VPS project would need to work closely with the Scottish Longitudinal Study.[83]

### 5.7. Existing linkage solutions

Lastly, it is important to consider what record linkage solutions already exist. As was mentioned at the beginning of the section, record linkage is a technique that is applied in several subject domains. Consequently, it is not surprising that a number of commercial packages exist aimed at a solution to the problem.[84] Whilst at this stage little more than a cursory investigation of these have been made, it would appear that none of these would be suitable to the needs of the VPS, either because of their inability to handle the complexities posed by historical sources, or due to the inability to handle variant multiple sources. However, one package that might prove of interest is the Freely Extensible Biomedical Record Linkage (Febrl) software, developed by the Australian National University.[85] This probabilistic linking software, which is available as OpenSource, is currently being evaluated for use by the Historical Census Project at the Minnesota Population Center, and it is recommended that any future VPS project take account of their findings.

## 6. Cross-comparability with other longitudinal studies

### 6.1. Background

Comparability with other data resources would be an important component of any future VPS. Ideally the VPS should be able to be analysed in conjunction with other longitudinal studies, both historical and contemporary, national and international. Where possible cross-comparative research would be possible, enabling results and finds to be placed in the context of long-term change and an international dimension.

---

[83] See http://www.lscs.ac.uk.

[84] Examples include Ascential Software, Trillium and Innovative Systems which are mainly orientated toward marketing applications, Search Software America, which is more generic, Linkage Wiz which is mainly health studies based.

[85] http://datamining.anu.edu.au/software/febrl/febrldoc/.

Because of these desirable goals for a VPS, it is important to undertake an investigation of other related longitudinal data resources. For two principal reasons: first, to see if it would be possible to learn from the experience of other national longitudinal database projects, especially those undertaking historical record linkage; second, to try and establish what measures might need to be taken in order to facilitate comparative research to be undertaken. What follows is an initial survey of other related longitudinal studies.

## 6.2. International historical longitudinal studies

### *6.2.1. Historical Sample of the Population of the Netherlands (HSN)*

The HSN is a representative longitudinal sample of about 80,000 people born in the Netherlands during the period 1812-1922, covering topics including age at marriage, religious affiliation, the number of children born, occupation, birthplace, literacy, social network and migration history.[86] The basic source material for the HSN consists mainly of the certificates of birth, marriage and death and of the population registers, which provide a continuous registration of the composition of households and the whereabouts of each individual. Of course, there is no complimentary source as the Dutch population register for the UK. The stated objectives of the HSN are four fold:

- to provide a representative dataset with which research can be done into social developments in the nineteenth and twentieth centuries;
- to provide a control group or groups that researchers can compare with their own research population;
- to develop the expertise that individual researchers usually cannot acquire in the limited time at their disposal;
- to offer the possibility for researchers to use the existing HSN dataset as base for their own research projects.

The HSN started in around 1987 with the formation of a working group with individuals drawn from several Dutch universities, across a range of social science disciplines. In 1989 this group was formally structured as a foundation with independent legal status. The new foundation set up an agreement with the International Institute of Social History in Amsterdam, to house their office and support it as far as the Institute could. This development encouraged the Ministry of Education and Research to give financial support for a pilot project. The province of Utrecht was selected because of its central position and because it had an appropriate number of births for a pilot project, about five per cent of the total of Dutch births between 1812 and 1920. Moreover, it is a region that

[86] See http://www.iisg.nl/~hsn/index.html. Much of this section is based on discussion carried out with Peter Doorn and Kees Mandemakers, who also kindly provided a number of published reports and internal documents.

is, in such aspects as religion, probably the most representative of the Netherlands as a whole. The data from the Utrecht pilot project were made available for academic research in 1994[87] and since then a subsidy of the Dutch foundation of scientific research has enabled the HSN to extend the original sample to cover the whole of The Netherlands. This second phase also extended the sources being used to include death certificates of those dying before 1940, marriage certificates, and population register entries, the last being particularly important since, as discussed below, these provide a 'ready-made' longitudinal perspective and add an otherwise missing household dimension.

The sampling of the HSN is based on tables with the official number of births for each year and each municipality. As a rule the whole period under investigation (1812-1922) is stratified into cohorts of ten years (except the first cohort, which contains the eleven-year period 1812-1822). These periods correspond to the administrative order of the certificates itself. Second, the sample is stratified according to geographic area of residence.[88] In the pilot Utrecht study a standard sampling frame of 1:200 births was applied, but led to rather unequal numbers per cohort surviving to the age of twenty due to differential mortality levels over time. As a consequence differential sample ratios were applied for different period, as follows: 1812-1872; 0.75 per cent; 1873-1902; 0.50 per cent; 1903-1922; 0.25 per cent.

The Netherlands is one of the few countries in the world that has kept a continuous population register starting as early as the mid-nineteenth century. In the early registers each household was entered on a double page, with the head of the household first, followed by his wife, children, other relatives, and other members of the household. Date and place of birth, relation to the head of the household, sex, marital status, occupation, and religion were recorded for each individual. All changes occurring in the household were recorded in the register.

Population registers remained in use until 1910 or 1920, after which a new form of continuous registration was introduced, consisting of single sheets, so-called family cards. The registration unit was no longer the household but the family. In the late 1930s, the population register was replaced by the personal card; and from this period the individual person became the registration unit in all municipalities. Since then the population register in each municipality consists of a collection of personal cards, containing nearly the same information as the population register. All persons who were alive in 1939 or were born after that year received a personal card. At the time of death, this card is removed from the files and sent to the Central Bureau of Statistics, where the

[87] See Mandemakers and Boonstra, *De levensloop van de Utrechtse bevolking*.

[88] In the case of the HSN this sampling method created a problem since in some urban districts the number of records new births was lower than the actual number of birth certificates as a result of people from the countryside going into the towns to give birth. This required the sampling frame to be enlarged in certain cases.

data on the card are used for statistical purposes; and then it is sent to the Central Genealogical Bureau. Copies of the cards are made available to the public, albeit without information that might infringe on privacy. The HSN makes use of the publicly released copies of personal cards for the construction of the dataset.

#### 6.2.1.1. Life courses in context

This is a collaborative project between the International Institute of Social Science History (hosts of the HSN) and researchers at the Netherlands Institute for Scientific Information Services and is aimed at supplementing the HSN micro-data with aggregate data from the Dutch census returns.[89] The project essentially proposes to digitise the published population and occupational census reports of the Netherlands for the period 1859 to 1947 and then link the tabulated data at the level of the municipality to the individual records of the HSN. This will allow the HSN database to be analysed in the context of the area level data. This is directly relevant of the proposed VPS, since one recommendation of this study is that a similar exercise is undertaken in conjunction with the VPS.

#### 6.2.1.2. Use and access to the HSN

The HSN has several products that can be ordered or downloaded. A demonstration version consisting of 100 records can be freely downloaded. This is meant as no more than an introduction to the HSN and encompasses five DbaseIV files, documentation on the database structure and a codebook, containing all the variable and value codes. Access to the full database has to be ordered from the HSN team, for which a handling charge may be made. Since the database contains information on living persons, the release data are anonymised and users are required to sign a licence agreement stating that they will not attempt to discover the identity of individuals within the database, or disclose information on individuals from it. The licence agreement also includes standards clauses on citation, publications and third-party access.

Interestingly, the HSN project team positions itself not just as a data producer and supplier, but also as a research collaborator. It presents itself as a service supporting research using the HSN, actively seeking collaborations and offering methodological support and the provision of bespoke in-house software. In parallel with this, the HSN team offer to create (given appropriate funding) augmentations and additional sub-samples linked to the main HSN, in effect producing what might be seen as booster samples for particular focus groups. This is done with the proviso that any new data must be offered to HSN for addition to the main database.

Research using the HSN to date has included:

[89] See http://www.lifecoursesincontext.nl/.

- Migration in Utrecht
- Reduced fecundity
- Regional differences in demographic behaviour
- Index of death certificates
- Geographical and social mobility in Zeeland
- Textile industry workers in Twente
- Germans in Utrecht in the 19th century
- Religious differences in infant and childhood mortality in the Hague
- Settlement of immigrants and their descendants, 1853-1960
- Family formation and living strategies
- Mobility of the Amsterdam poor
- Early-life conditions, social mobility and longevity
- European migration to the Dutch East Indies

### *6.2.2. Historical Census Project, Minnesota Population*

The Minnesota Population Center has performed a major task through its Integrated Public Use Microdata Series (IPUMS) programme to create a number of national census samples covering all fifteen US federal censuses from 1850-2000. A particular achievement of this long-term project has been the harmonisation and standardisation of the various sample census datasets over time, producing uniform codes were possible and integrated, coherent documentation. This exercise has greatly facilitated the use of the national census samples to study change over time.[90] The IPUMS team have also devoted a lot of effort to produce harmonised coding schemes across census data from different countries. This includes an historical project using the 1881 British census, in which the University of Essex are the UK partner.[91].

The IPUMS team have also embarked on a project to create linked representative samples of individuals and family groups from the existing national samples of the censuses of 1860, 1870, 1900, and 1910 to the 1880 census, which, as in the case of Great Britain, exists in computerised form at the 100 per cent level. Using record-linkage techniques it is proposed to construct linked samples covering pairs of census years: 1860-1880, 1870-1880, 1880-1900, and 1880-1910. However, unlike most linking projects which focus on maximising the number of accurate links, this project will be designed to maximise the representativeness of the linked cases, paying close attention to sources of selection bias, and ignoring much of the information routinely used by other record linkage procedures. Thus, for linking purposes the project uses only those variables which are deemed to be 'fixed' over time: surname, forename, age (as in year of birth), birthplace and gender. It ignores variables that

[90] See http://www.ipums.org/usa/index.html.
[91] See http://www.nappdata.org.

can change over time: occupation, marital condition, place of residence. This also means that females are treated differentially for linking purposes since their surname is at risk of change as a result of marriage. Thus the project plans three linked samples: all males; females who do not marry between a census interval; married couples. One key difference is that the Minnesota project does not intend to use civil registration data to link in combination with the census records, which means that they cannot observe female name changes through marriage.

Although none of the three linked sample groups is representative of the entire population, the project's goal will be to make each category representative of its defined universe. It is intended that the male individual sample will be general purpose, useful for studying economic and geographic mobility, transitions to adulthood, changes in family composition, and retirement. The female sample will be useful for studying many of the same topics, but will apply to the subset of women who do not change their surname between censuses, and therefore will be inappropriate for some topics. The married-couple samples will offer the greatest reliability, since it will allow linking on characteristics of both husband and wife, and will be especially useful for topics relating to fertility, child mortality, and age of leaving home. Because it is restricted to the continuously married population, however, it will be less useful for population-wide generalisations about social and geographic mobility. Although the linking 'unit' will be individuals or couples, information will also be captured on the characteristics of all other co resident household members.

Once the project has established the links, it will add the omitted 1880 variables covering health, education, unemployment, and other characteristics for all persons residing in a household containing linked individuals. There is insufficient information at present to estimate the total size of the linked samples, but it is estimated that approximately 600,000 cases will be created across all linked samples. The project will also provide a confidence score for each linked case, indicating the strength of the similarity measures between the two cases, and construct weights for the linked samples that maximise the representativeness of the linked cases with respect to occupation, birthplace, parental birthplaces, age, sex, race, marital status, literacy, state of residence, urban residence, and school attendance.

Although MPC differs in approach and outcomes from that proposed for VPS, it will undoubtedly provide opportunities for comparative research on nineteenth-century American and British populations. The software that is being produced for automatic linking by the Minnesota team could also have implications for the VPS. The VPS team are in close contact with the Minnesota project and they have indicated their willingness to co-operate fully, sharing experience and outputs, such as linking software, coding dictionaries and related look-up tables.

#### 6.2.2.1. Use and access to the IPUMS data

All of the census data samples generated at Minnesota are freely available on-line for academic research. Users are asked to acknowledge and cite the data sources and provide copies of any resulting publications. Web-based interfaces to the data have been developed that allow users to design bespoke datasets, specifying subsets by year and place, and then selecting the variables they wish to include within their dataset. A query is then run and the required dataset is generated automatically ready for the user to download.[92] It is not yet known how the linked longitudinal data will be disseminated, but it is expected that it will be released in a similar on-line system.

The cross-sectional census sample data for the US has been very widely used. The IPUMS on-line bibliography lists 1,806 published research outputs, many of which are of an historical nature.[93]

### *6.2.3. Swedish historical longitudinal studies*

Sweden has given rise to a number of historical longitudinal studies. In part, this is an outcome of the quality of Swedish historical demographic sources. However, curiously, these tend to be local studies, with little or no integration between them.

#### 6.2.3.1. The Demographic Data Base (DDB)

The DDB consists of linked entries for individuals recorded in nineteenth-century parish registers, giving information on births, marriages and deaths, together with catechetical lists and migration lists.[94] Initially seven individual parishes were chosen, representing a variety of industrial and social environments. However, due to linkage attrition caused by population mobility, resulting in a significant loss of complete life-histories the DDB switched to linking individuals across whole regions. So far, two regions (Skellefteå and Sundsvall) have been completed and work has begun on a third (the town of Linköping, together with thirty-five surrounding parishes). Linkage is carried out via a mixture of automated and manual linking.

Researchers wishing to use the data are generally encouraged to visit and work with the project team in Umeå, although specific files can be written on a case by case basis. The DDB team are enthusiastic about sharing their knowledge and experience, especially in record linkage.

---

[92] See http://www.ipums.org/cgi-bin/usa/extract_main.pl?process=IntroRevise for the IPUMS data extraction system. This is very similar to the system designed by UKDA/AHDS History for disseminating the 1881 and 1851 census data. See <http://ahds.ac.uk/history/collections/chccaccess.htm>.

[93] See http://www.ipums.org/usa/research.php.

[94] See http://www.ddb.umu.se/index_eng.html.

#### 6.2.3.2. The Scanian Demographic Database (SDD)

The SDD is based at the Department for Economic History at the University of Lund, and is constructed from parish records on vital events, migration registers, catechetical examination registers and poll-tax registers for ten parishes in the Scania region between the middle of the 17th century and 1900.[95] Use of the database is primarily restricted to researchers based in Lund and their associates, however, it has been used extensively to investigate a wide range of topics, including: mortality, fertility, marriage, migration, age at leaving home, widowhood, domestic service, ageing and retirement, inheritance and social mobility.[96]

#### 6.2.3.3. The Stockholm Historical Database (SHD)

The SHD project began in the mid-1970s, with the aim to create a complete demographic database of all residents of Stockholm, based on the special population register of the Roteman Archive, covering the period 1878 to 1926. Once completed, it is estimated that the linked database will contain information on some 443,000 individuals. Unlike many historical longitudinal databases, which concentrate on demographic features, due to the quality and richness of the Roteman Archive, the database contains additional information on poverty, medical provision, crime, education, and family and household structure. Another interesting feature of the SHD is that the underlying individual and household records have been linked to an historic Geographical Information System of Stockholm allowing the data to be mapped and visualised spatially.

Unlike the proposed VPS, SHD is not a sample as such, but rather an attempt to track every individual resident in Stockholm over the period in question. Equally, record linkage is not a major issue for the project since the underlying population register on which it is based is a continuous longitudinal data source.

Data are made available to users via CD/DVDs upon request.

### *6.2.4. Programme de recherché en démographie historique*

The Programme de Recherché en Démographie Historique (PRDH) is a project of long-standing based at the University of Montreal that aims to reconstitute the population of European descent in the St. Lawrence (Quebec) for the seventeenth and eighteenth centuries.[97] The main sources used are a series of high quality baptismal, marriage and burial certificates. The database integrates all the entries of marital status prior to 1800 in the parochial registers as well as

[95] Olsson and Reuterswärd, *Skånes demografiska database*.

[96] The document at http://www.ekh.lu.se/ed/papers/_vti_cnf/SDD%20publications.pdf lists some 87 publications using the SDD.

[97] See http://www.genealogie.umontreal.ca/en/lePRDH.htm.

the biographical files of the individuals who appear there. Each file identifies the individual by name, specifies the dates and places of its birth, of its death and of its marriages (if they occurred) and connects it to her/his parents, to children and to spouse, allowing the inter-generational analysis. Each personal file specifies the date and place of birth, as well as details of marriages and deaths, and the subsidiary and marital bonds maintained with other individuals. The basic information is supplemented by socio-demographic characteristics such as occupation, literacy, place of residence and, in the case of immigrants, the place of origin, as drawn from the historic source records.

The record-linkage was achieved through a mixture of automated and manual approaches. Unambiguous links were resolved automatically, while others are resolved manually, although the exact methodology remains unclear.

Interestingly, the PRDH database has been used not only to address issues related to historical demography, but also for research in anthropology, human biology and genetics. The on-line publications database lists over 200 research outputs.[98] It would seem that users gain access to the data by going to Montreal and working on them there in collaboration with the PRDH team.

### *6.2.5. Other historical studies*

A project based at the Norwegian Historical Data Centre, University of Tromso, is attempting the linkage of censuses for the period 1865-1910, together with selected church records for sample of Norwegian parish.[99] The *Aranjuez Database*, created by the Population and Society Research Group (GEPS), Madrid, is a multi-source linkage exercise for the town of Aranjuez, fifty kilometres south of Madrid, linking padrones (census type listings) for 1877, 1905, 1912, 1945, 1960, 1975 with various church records and military records.[100] The *BALSAC* population register project based at the University of Quebec, links church records for Quebec from seventeenth to twentieth centuries.[101] The *Base TRA Patrimoine* is a French, national level, multi-source longitudinal study, compiled with the help of family historians and genealogists, for the period 1800-1900, being a sample of individual's whose surname starts with either the letter T, R or A. The *Founders and Survivors: Tasmanian life courses in historical context* project is based at the School of Historical Studies, University of Tasmania, and is a longitudinal study of Australian settlement from 1803 to 1856 based on convict records, census material, Inquests, criminal trials and immigration and inter-colonial migration records. Interestingly the study is being used to compare with contemporary longitudinal health studies. The *Geneva Database* project studies the population of Geneva 1816-

[98] See http://www.genealogie.umontreal.ca/en/Bibliographie.htm.
[99] See http://www.rhd.uit.no/.
[100] See http://www.ucm.es/info/geps.
[101] See http://www.uqac.ca/balsac/.

1850 using census and registration records, sampling all persons with whose surname starts with B, plus all household members from the census where the household head's surname starts with B.[102] Multiple sources including tax registers, cadastre and military conscription lists, plus population registers have been used for the *Historical demography of the Liege region* project, which focuses on the town of Verviers for the period 1806-1900.[103] Again for Verviers, those whole whose surnames starts with the letter B are sampled. The *Koori Health Research Database* (KHRD), based at the School of Population Health, University of Melborne, is a study of the Aboriginal population of Victoria, 1840-1930, using street directories, police censuses and records, prison records, school registers, asylum registers, army records, records of the Aboriginal Protection Broad and Mission Station records. It is associated with the *Melborne Lying-in Hospital Cohort: 1857-1900* project which is also a multi-source linkage exercise based on the Midwifery registers of the lying-in hospital 1857-1900. It also uses Charity Organisation and Ladies Benevolent Society records.

## 6.3. Present day longitudinal studies

In addition to ensuring that a VPS is made comparable, where possible, with other international historical longitudinal studies, it is important that it also maximises the potential comparability of the VPS dataset with major contemporary longitudinal surveys. Secondary analysts, whether historians or social scientists interested in an historical perspective, may wish to make comparisons between the latter half of the nineteenth century covered by the VPS and the latter part of the twentieth century (to present) covered by the major contemporary longitudinal surveys. Changes in household and occupational structure appear to be the two most likely areas of comparative investigation. Given the increasing volume of machine-readable aggregated nineteenth-century census, civil registration and other data, aggregate comparisons between the late nineteenth and late twentieth centuries are already possible for a number of variables. The individual-level data of the proposed VPS sample and contemporary longitudinal surveys should permit more detailed and sensitive comparisons to be made. For example, one could examine trends in age at first marriage by occupational group or country of parental birth. Analysis of a VPS alongside the British Household Panel Survey would permit a comparison of the intergenerational dynamics of occupational/class mobility a century or more apart using the detailed household and intergenerational information present in both these sources.

[102] See http://www.unige.ch/ses/istec/.
[103] See Alter, Family and the female life course.

In order to estimate the potential for comparative research, a variable-by-variable comparison of the expected content of the proposed VPS sample with major contemporary longitudinal surveys. For each variable, the aim was to identify whether such a comparison was possible and if so whether secondary analysts would benefit from additional guidance/data preparation work (carried out by the VPS project) to perform such a comparison. The contemporary longitudinal surveys examined were:[104]

- the Office for National Statistics Longitudinal Study (1971 onwards, ONS LS);
- the National Child Development Study (1958 onwards, NCDS);
- the 1970 British Cohort Survey (1970 onwards, BCS);
- the Millennium Cohort Study (2000 onwards, MCS);
- the British Household Panel Survey (1990 onwards, BHPS).

All these surveys are freely available to the research community. The ONS Longitudinal Study is available through the Centre for Longitudinal Study Information and User Support (CeLSIUS).[105] The three birth cohort surveys and the BHPS are distributed by the Economic and Social Data Service (ESDS).[106]

### *6.3.1. Comparability of VPS variables*

Table 6.1 compares the variables/measures anticipated to be in the VPS sample with analogous variables (where these exist) in the major contemporary longitudinal surveys. For most variables/measures the ability (or inability) to compare is clear cut and depends upon the presence (or absence) of an analogous variable in at least one of the contemporary surveys (or, more specifically, the anonymysed versions of these surveys that are made available to secondary analysts). Occupation and social class proved more complex issues, some form of comparison is possible, but might prove beyond the knowledge base or resource constraints of some secondary analysts. As a result there is an argument that a future VPS project should conduct such work to facilitate such analysis. This work entails coding the VPS occupation/socio-economic status data to a variety of occupational and socio-economic status schema to allow forward comparison with the contemporary surveys and also ensure consistent comparison within the span of the VPS (and allow for subsequent extension of the VPS to 1911). This work is detailed in section 6.4.

[104] The earliest of the contemporary ongoing longitudinal surveys, The National Survey of Health and Development (a cohort survey of respondents born in 1946) was not included here because precise information on the content of this data resource is not in the public domain.

[105] See http://www.celsius.lshtm.ac.uk/.

[106] See http://www.esds.ac.uk/.

Tab. 6.1: Comparability of VPS data with contemporary longitudinal surveys

| Variable/measure | In contemporary longitudinal surveys? | Compatibility of VPS with surveys listed in previous column |
|---|---|---|
| *Name data* | | |
| Surname | No. Data always anonymised before distribution to secondary analysts. | Not possible. |
| first name | No. Data always anonymised before distribution to secondary analysis. | Not possible. |
| *Demographic data* | | |
| Age (reported) | In all. | Yes[1] |
| Sex | In all. | Yes. |
| Marital status | In all. | Yes[2] |
| Fertility history | BHPS (indirectly once span is long enough, and retrospectively in Wave 2), NCDS, BCS and MCS (only full history for NCDS, given respective age of BCS and MCS); ONS LS (inc. pre 1971 fertility history) | Yes, though will only be known post- 1851 in the VPS. |
| Mortality | Only partial (since duration of contemporary studies are all less than a full lifespan). | Not for most purposes (for reason given in left-hand column)[3] |
| *Family/household structure data* | | |
| Relationship to head of household | BHPS, ONS LS. | Yes. |
| Household structure | BHPS (full information for every household member); ONS LS, NCDS, BCS and MCS have summary household structure information. | Yes[4] |
| *Geographical data* | | |
| Place/area of residence | No. Location of residence is only generally available at a very broad scale (typically Government Office Region). Special/restricted access may be possible to gain more detailed spatial information. | No. |

Notes to Table 6.1:

[1] Age 'heaping' in the VPS should be considered if very detailed age comparisons are made.

[2] The rarity (and effective absence of recording) of cohabitation, separation and divorce in the historical census data should be noted.

[3] Note also the inevitable (slight) deficiency in contemporary surveys as well as the more substantial deficiencies of the VPS data in terms of the inability to delineate between attrition, emigration and mortality.

[4] The VPS household definition will be based upon contemporary enumerator definitions and will include lodgers and other individuals who would not be defined as household members in contemporary surveys. In more technical terms, contemporary surveys employ a 'housekeeping unit' definition as opposed to the 'household-dwelling' definition of historical censuses, but the VPS relationship could be coded to facilitate comparisons to be made.

Table 6.1 includes variables in the VPS resulting more or less directly from transcription of the enumerators' returns (or civil registration data) and those that will be derived by the VPS (coded occupation fertility history).

## 6.4. Recommendations for coding occupations/social status

As made clear in Table 6.1, comparability between the main VPS sample and the major contemporary longitudinal surveys is generally a straight forward issue: comparison is generally either impossible or possible and relatively simple. The variable presenting greatest complexity is coded occupation, and by extension social status. In order to deal with this complexity the VPS will need to consider coding occupational data to a number of classifications.

In coding occupations within a project such as the VPS, three major issues need to be addressed. First, it will be important to code occupational titles in such a way that the resulting database can be analyzed alongside contemporary longitudinal data without the user having to perform major recoding and occupational mapping exercises. Second, the same can be said for analyzing the VPS in conjunction with other major historical longitudinal data for other countries. Third, in order both to contextualize the data and draw comparisons with other published nineteenth-century statistical sources, it will also be desirable to code occupations in such a way that they can be compared with the various published tables of the Registrar Generals over the period in question. It would be impossible to satisfy all of these three requirements within a single code, especially since researchers will require social class to be classified alongside occupations. Thus a multi-dimensional occupation coding exercise is the only practical way forward.

### *6.4.1. Comparability with contemporary data*

In order to achieve the first of these goals, it will be important for the VPS to produce a variant of the Standard Occupational Classification 2000 (SCO2000) which is a revision and replacement of the earlier Standard Occupational Classification 1990 (SCO1990).[107] The SCO2000 classification scheme is now being applied to all major government surveys, although ideally the VPS should produce both SCO2000 and SCO1990 codes for the purposes of comparability since earlier survey (mainly pre-2001) will be coded to SOC1990. Likewise, the 1990 Standard Industrial Classification (SIC) should be used to facilitate cross-comparability, together with the International Standard Classification of Occupations (ISCO-88)[108] and the European Union variant of this, ISCO 88 (COM).[109] Whilst this sounds like a tall order, most of these contemporary coding schemes are inter-related and mappings between them exist in various forms exist.[110] Software tools have also been developed to assist with coding of SOC and SIC classifications.[111]

In addition to these classifications, it will be important for the VPS to classify occupations according to social class. Traditionally, the most used scheme for classifying social class has been the Registrar General's Social Class categories, originally based on the analysis of fertility undertaken in conjunction with the 1911 census.[112] In 1994 the then Office of Population Censuses and Surveys, now part of ONS, commissioned the ESRC to undertake a review of its social class classifications. This process resulted in the recommendation of the National Statistics Socio-economic Classification (NS-SEC).[113] The VPS should clearly endeavour to map to this new standard, together with the so-called 'Goldthorpe Schema' from which it is in part derived and which is widely used by social science researchers.[114]

The major problem for the VPS in producing these various classifications is that SOC2000 and NS-SEC ideally require information on employment status

[107] ONS, Standard occupational classification 2000. Volume 1; ONS, Standard occupational classification 2000. Volume 2.

[108] See http://www.ilo.org/public/english/bureau/stat/class/isco.htm and International Labour Organisation, *International Classification of Occupation, 1988*, (Geneva, 1990).

[109] See http://www.warwick.ac.uk/ier/isco/brit/btext1.html.

[110] For example, ONS, *Standard occupational classification 2000. Volume 2*, includes an alphabetical list of some 26,000 mapping SCO1990 to SOC2000. The BHPS team at the University of Essex have also undertaken extensive mapping exercises in order to provide multiple codes for the BHPS data. In addition to using the SOC2000 and SOC1990 schemes the BHPS is coded to the four digit level for ISCO-88 and to the 'activity' level for SIC.

[111] For example, Cascot, developed by the Institute for Employment Research at the University of Warwick, which is also used by BHPS. See http://www.warwick.ac.uk/go/cascot.

[112] Szreter, 'The genesis of the Registrar-General's social classification'.

[113] Rose, Pevalin and O'Reilly, *The NS-SEC*; Pevalin and Rose, *A researcher's guide*; O'Reilly and Rose, eds, *Constructing classes*.

[114] Goldthorpe, *Social mobility*; Goldthorpe, 'The 'Goldthorpe' social class schema'.

(whether an employer, self-employed or employee; whether a supervisor), the nature of the business of the employer (including number of employees in employing organisation), qualifications and membership professional bodies. While the historic British censuses contain vague references to employment status (employers where asked to identify themselves and the numbers of persons employed, equally terms such as 'Master' or 'Apprentice' were used) from 1851, explicit questions on employment status ('Employer'; 'Employed' and working 'On own account') were only included from 1891 onwards. Routine questions on the nature of the business of the employer, number of employees and qualifications were not asked until after the period considered by the VPS. Thus, it will only be possible to map to SOC2000 and NS-SEC approximately, yet still in a form robust enough to facilitate comparative analysis.

### *6.4.2. Comparability with historical data*

There is no single occupational classification scheme that is used by historians which acts as what might be termed a standard, however, the so-called 'Booth-Armstrong' code has been used relatively extensively by those working on British historical census information.[115] At an international level an interesting development has been the creation of the Historical International Standard Classification of Occupations (HISCO).[116] This has been developed by a team of researchers drawn from across Belgium, Canada, France, Germany, the Netherlands, Norway, Sweden and UK and is based on ISCO68, the International Labour Organisation (then know as Office)'s 1968 International Standard Classification of Occupations.[117]

Of potentially greater importance for the VPS, however, if only because the 1.5 or so million occupational tiles recorded in the 1881 British censuses have already been mapped to it, is the occupational classification scheme being used by the North Atlantic Population Project (NAPP) which harmonizes raw occupation strings across five complete country censuses (and four languages).[118] The NAPP occupation code has been adapted from the HISCO and like the ISCO and SCO families of codes is hierarchical, in the sense that each digit in the five digit codes introduces a new level of detail. Codes sharing the same first 1, 2 or 3 digits are considered to be increasingly similar. For example, all people working in agriculture have the first digit 6. The first digit of a code indicates the 'Major Group' an individual's occupation falls into. The second digit indicates a 'Minor Group' distinction. Continuing the previous example,

[115] Armstrong, 'The use of information about occupations'. See also the extensive analysis of coding schemes used in conjunction with historical British census data in Mills and Schürer, 'Employment and occupations'.
[116] Van Leeuwen, Maas and Miles, *HISCO*.
[117] International Labour Office, International standard classification of occupation.
[118] See section 6.2.2 and Roberts, *et al*, 'Occupational classification'.

people who have the first two digits '61' are farmers – who may specify what they are cultivating or tending – and farm managers. Thus, as well as sharing the characteristic of working in agriculture (6) they also share the characteristic of being owners or managers (61). The first 3 digits denote the 'Unit Group' of an occupation. At the third digit level, more detail is introduced, for example, the unit group '612' indicates 'Specialized Farmers'. Within this unit group, 4th and 5th digit distinctions, known as 'titles' or 'headings' are made. For example, 61220 indicates 'Field crop farmers', and 61230 indicates 'Orchard keepers and fruit farmers'. Thus what might be termed general occupational titles, lacking specific detail, such as 'Metal smelter' or 'Furnaceman', will end with the last two digits of a code being '00'.[119] Headings ending in digits other than '0' are generally reserved for frequent responses specific to a particular country that probably belong with responses sharing the same first four digits of the heading. For example, 61115 (Husbandman or cottar) and 61117 could both be classified with other general farmers (61110), but occurred frequently enough in Norway and Canada respectively that we felt they should be given a separate code for easy identification.

In comparison to the HISCO coding scheme, the NAPP project has reduced the overall number of headings, while still introducing new ones, and retaining more detail from vaguely specified occupations. In general, the fewest number of changes to the structure of HISCO is with major groups 0 and 1 (professional workers). The codes within major groups 2 (administrative and managerial workers), 3 (clerical workers), 4 (sales workers), 5 (service workers), and 6 (workers in agriculture) had been re-organised in various ways, while the most substantial revisions, including moving occupations between major groups occurs with the manufacturing and transport major groups (7, 8, 9). In particular, codes have been created for vague responses in the form 'Works in [specified type of] factory' or 'Works in cotton mill'.

NAPP have also eliminated codes where HISCO made distinctions that were not consistently made in nineteenth-century census data. For example, HISCO distinguishes between hand and machine spinners. However, in nineteenth-century census data most spinners did not specify whether they were spinning by hand or machine, more often giving information on what they were spinning (e.g. cotton, wool, silk). Thus, the NAPP project coding scheme can be seen as offering a finer level of granularity, and also taking greater account of the nature of nineteenth-century census sources. Given this, and the existing work that has already gone into coding the 1881 census in NAPP, it is recommended that a future VPS should seriously consider using NAPP as the base occupational code from which other coding schemes are derived.[120]

[119] In these cases the code is 72100 – 721 being the unit group 'Metal smelter and furnace workers'.

[120] A mapping between NAPP and HISCO already exists. Equally, mapping exists between HISCO and ISCO68 (see http://historyofwork.iisg.nl/detail_page.php?act_id=35200) and

### *6.4.3. Comparability with nineteenth-century classifications*

Unfortunately, although a number of similarities exist, the Census Offices of the nineteenth-century produced a different occupational classification scheme for each census year in the period of the VPS.[121] Thus, for the researcher wishing to compare results from any one of the published Census Reports for the period, this could present a problem. However, work has been carried out to produce a single code, in the form of a mapping exercise, covering all the Census Office occupational classifications from 1851 to 1921.[122] The extension of this scheme to the twentieth century is useful in that provision can be made to re-classify occupations according to the original social class categories introduced by the Registrar General in conjunction with the census of 1911, and the modifications to the social classification introduced in 1921. Thus, it is recommended that this existing work be augmented and adapted for incorporation into the VPS.

## 6.5. Database design and dissemination issues

Although it is premature at this stage to envisage exactly in what format a completed VPS might be disseminated to future researchers, the VPS project undertook some preliminary investigations in this area. Regarding existing historical longitudinal studies, it is the case that no standards for data dissemination currently persist. Of the historical longitudinal studies investigated by the VPS project, data are produced in a variety of structures and formats, with a number of proprietary software packages being used to support them. Of greater relevance to the VPS, therefore, is the experience of the BHPS, in part because the VPS should aspire to producing a data distribution format that would ease and facilitate cross-comparison with the BHPS, and in part because the BHPS can in many ways be seem as an exemplar model.

The BHPS database has been specifically designed to aid 'matching', 'aggregating' and 'distributing' information across the database.[123] Matching allows information to be compared for the same individual across different files, for example, the variable type but in different waves. Aggregating allows information from one level to be calculated and related to a different level, for example, calculating total household size from the individual level records and writing this to the household level. Distributing is similar to aggregating in that it associates information from one level with records at another level, for ex-

translation tables exist between ISCO68 and ISCO88, and from ISCO88 to SOC90 (see http://www.cf.ac.uk/socsi/CAMSIS/occunits/distribution.html).

[121] Differences also exist between Scotland and England and Wales.

[122] This has been described as the 'Cambridge Code' and is provided in Schürer, 'Understanding and coding the occupations of the past'. See also the important discussion of this problem in Woollard, 'Conceptions of work'.

[123] See Taylor, *BHPS user guide*.

ample, writing area level mortality and fertility rates to individual level records. Although the BHPS obviously contains a far greater volume of information than would a VPS, in theory there is no reason why the VPS should be able to mimic the basic BHPS data structure, thus allowing VPS records to be disseminated in a near identical form to those of the BHPS.

Other lessons can also be learnt from the BHPS. In particular, there will be significant advances to creating sampler datasets, as ESDS has done for the BHPS for given themes.[124] Available for tabulation online, these have been particularly successfully in the use of teaching, and provide an easy entry point into what is potentially a highly complex and confusing database. Thus also points to the need for clear and concise user guides and related documentation. Again the BHPS provides a model of best practise, with excellent online and paper-based documentation. Indeed, the need for adequate documentation is possibly even greater in the case of the VPS since social and economic historians (although they would not be the only user community) are generally less experienced than other social scientists in the complexities associated with the analysis of large longitudinal databases. Thus, adequate support mechanisms also need to be put in place, via specialist 'how to' guides, training workshops and sessions.[125]

# 7. Potential of Relating other materials to the VPS

## 7.1. Background

From a research point of view it may prove desirable to link (implicitly or explicitly) records from the VPS to other economic and social historical materials. In consequence a number of potential sources were assessed, drawing on input from the user consultation. In considering materials that might be used in conjunction with a VPS these were split into three broad categories:

- individual level;
- household level;
- area level.

Whilst it is recognised that these categories may overlap in some cases, this broad division was deemed helpful in identifying those materials that might be explicitly linked to the panel study, at either the nominal or household level, as opposed to materials that might be linked implicitly to a higher level spatial unit in order to provide summary contextual information that could be utilised

[124] See http://www.data-archive.ac.uk/findingData/bhpsTitles.asp.

[125] These concerns of adequate documentation and training for users were recently strongly echoed at a workshop on 'Disseminating and analyzing longitudinal historical data', held at the International Institute of Social History, Amsterdam (March, 2006). See Alter, Gutmann and Mandemakers, 'Problems and possibilities'.

in analysing the data. Such information might be organised at the level of the census enumeration district, the parish, the registration district, the county or an appropriate regional unit. These are collectively termed 'community' level materials.

## 7.2. Individual level

Ideally, the optimum situation would be to have individual level materials of interest that were systematically created and exist routinely for each individual within the population. However, for the nineteenth-century very few administrative records were created for every individual living in the country. Indeed, the census and civil registration were the only administrative processes that required every individual to be recorded, which, in turn, is why they are particularly relevant to the VPS. There are, however, a number of records which, although only available for a sub-set of the overall population, may be useful in enriching the panel for specified members.

### *7.2.1. Parish Registers*

Ecclesiastical parish registers, recording the enactment of the religious ceremonies of baptism, marriage and burial could be said to have formed the backbone to many studies in historical demography in the UK, especially England where the legislation for keeping parish registers was laid down in 1537. They have, indeed, often been linked over time, primarily for periods before the nineteenth-century, in order to generate 'family reconstitutions' from which age-specific fertility and mortality rates can be derived.[126] However, they have not tended to be used much for the post-1837 period when civil registration was introduced in England and Wales, in part because the various publications of the Registrar General provide a more useable source of information on demographic patterns from this date, but perhaps more importantly, parish registers are generally considered too incomplete from the early-nineteenth-century due to the general rise in non-conformity and religious dissent.[127]

In relation to the VPS, parish registers may provide two important opportunities. First to 'back fill' by supplying information on the demographic history of individuals within the base sample prior to 1851, in particular, helping to complete fertility and marriage histories. Second, in providing indicators to religious denomination, although the civil registers can also provide information on this topic.

[126] See Wrigley et al *English population history*.

[127] It has been suggested that the 'breakdown' of the parish register system may have been less marked in certain rural communities in the nineteenth century. See Hinde, 'The population of a Wiltshire village'.

The largest collection of parish register materials available in machine-readable form is the International Genealogical Index (IGI) compiled by the Church of Jesus Christ of Latter-day Saints (LDS).[128] This mainly consists of baptism and marriage records, and for England and Wales covers some 6 million entries for the period c.1537 to c.1860. Early investigations suggest that the LSD may be willing to make their entire GB parish register database available for academic research via the UKDA.

In addition to the LDS IGI , an index of the Scottish Old Parish Registers (OPR) has been compiled from the surviving registers of some 900 parishes of the established church (Church of Scotland) for the period 1553 to 1854. There are two main indexes, as follows. The births and baptisms index includes: full date of the event, surname, forename(s), gender, parents' names (but some mother's names are missing), parish in which the birth or christening was registered, and source reference. The banns and marriages index contains: full date of the event, the surname(s), forename(s), gender, the parish where the proclamation of the banns or the marriage was registered, and the source reference. There are currently no national name indexes for deaths/burials before 1855, although local indexes do exist. If necessary, GRO Scotland (GRO(S)), who hold the data would appear to make them available to a future VPS project.

### *7.2.2. Burial records*

Separate, but clearly related to parish registers are burial records. The main distinction being that in addition to the recording of the burial service (to be found in parish registers) additional records exist, especially from the nineteenth century onwards, of burials made by the cemetery where the body was actually buried (or burnt). The most comprehensive collection of burial records in computerised form for England and Wales is the National Burial Index (NBI). This is an on-going project which started in 1994 and is conducted by the Federation of Family History Societies (FFHS) whose members transcribe and computerise the burial records. A parallel project, has subsequently started in Scotland under the supervision of the Scottish Association for Family History Societies. The NBI is being compiled from multiple sources: parish registers, including, importantly, non-conformist registers; bishops' transcripts; and cemetery records. To date some 20 million records have been entered into the database. Coverage varies from county to county – Devon, Hampshire and Middlesex are particularly poor as their FHS are not yet taking part in the project – yet others are generally well covered. In many cases the post-1813 tends to be better covered than earlier periods because of ease of reading the registers, but in some smaller rural parishes the late nineteenth-century has yet to be

[128] See http://www.familysearch.org.

covered since the church registers are still in use and have not been deposited with local archives. Cardiganshire, Dorset, Essex, Leicestershire, Lincolnshire. Nottinghamshire, Rutland, Staffordshire, and Suffolk all appear to have a relative proportion completed to the end of the nineteenth-century. Other counties are rather mixed. The project continues to grow quite rapidly.

Other local-based projects have transcribed municipal burial records. One such project, the Kingston Local History Project, has successfully linked these to census data.[129]. There is probably potential for similar work to be done by local volunteers, especially where indexed transcriptions already exist.

### *7.2.3. Trade and street directories*

These list only a small proportion of the population, but can provide important changes occurring between census dates, as well as enhanced information about a person's occupation or trade.[130] For example, trade directories may be helpful in providing information on where the trade was carried out, as opposed to where an individual with a certain trade lived. Trade directories and Street Directories are available for all counties, most towns and cities, and for most years throughout the nineteenth and twentieth centuries.[131] Major national collections are at the Institute of Historical Research and at the Guildhall Library. Local collections are held at County Record Offices, large libraries and some Family History Societies. A major project based at the University of Leicester has digitised over 600 trade and street directories for the period 1750 to 1919, with most dating from the 1850s.[132] These are held primarily as digital images of the pages, but some have been converted into a searchable text database using optical character recognition software, and more could be done in the future.

### *7.2.4. Wills*

Although only a relative small proportion of the population made wills in the nineteenth century, they could provide a wealth of information for any individuals within the VPS who could be identified as will makers, including information that could not readily be gained from other sources, such as wealth, debt and ownership. The critical date for wills in England and Wales is 1858, when probate laws changed. From this date all 'proved' wills should be centrally registered, and indices to them exists, with copies for the 1858 to 1943

[129] French, Sullivan and Warren, *Burial Registers for Kingston upon Thames*.
[130] Trade directories have been linked successfully with census data, for example, in the case of Pooley, 'Residential mobility'.
[131] Shaw and Tippet, *British directories* and Mills, *Rural community history*.
[132] See the project website at http://www.historicaldirectories.org. The data have also been deposited with AHDS History at the UKDA and are currently being processed there.

period in England available from TNA and the Family Records Centre. Copies of wills proved in Wales between 1858 and 1940 are held in the National Library of Wales. Pre-1858 TNA has the records of the Prerogative Court of Canterbury, which cover mainly the southern half of the country, these have recently been digitised and are searchable by name.[133] Many wills relating to this period for the north of the country were proved at the Prerogative Court of York and are held at the Borthwick Institute of the University of York. [134] The records of the minor Probate Courts are deposited in county record offices or other local archives under a variety of indexing systems.

The legal system in Scotland regarding probate was rather different. The National Archives of Scotland holds all Scottish wills proved from the sixteenth century up to 1991. These are also searchable online before 1901[135] and if required could be made available to the VPS via the GRO(S).

### *7.2.5. Electoral registers and Poll Books*

Following the 1884 Electoral Reform Act an increasing proportion of the population had and exercised the right to vote. For the period before the introduction of the secret ballot in 1872 many poll books have survived and thereafter electoral registers were compiled annually. The largest national collections of poll books are held at the Guildhall Library and at the Institute of Historical Research. There are local collections in County Record Offices, some of whom also hold collections of electoral registers.[136] The largest single collection of electoral registers is held in the British Library's Official Publications and Social Sciences section, which totals some 20,000 registers for the period 1832 to 1931. Coverage is described as 'modest' to 1885 but 'good' from then until 1915.[137]

Unfortunately, relatively few nineteenth-century poll books and electoral registers exist in electronic form, thus linkage to the VPS would be a major project, but as these are printed sources in standardized format it could be a suitable task for volunteers, and in particular could provide valuable information on intra-census residential moves.[138]

---

[133] www.nationalarchives.gov.uk/documentsonline/wills.asp.
[134] See, as an example, Webb and Hemingway, 'Improving access'.
[135] See http://www.scottishdocuments.com.
[136] Sims, *A handlist of British parliamentary poll books*; Society of Genealogists, *Directories and poll books;* Gibson and Rogers, *Poll books c.1696-1872.*
[137] Cheffins, *Parliamentary constituencies and their registers since 1832;* Gibson and Rogers, *Electoral registers since 1832*.
[138] Dennis, 'Intercensal mobility'.

### *7.2.6. Miscellaneous*

For those individuals who were found to be resident in an institution in the VPS (and who are likely to be among the most difficult to link) the linkage of institutional admission and discharge records may be beneficial. Two particularly useful sources that are available electronically (but at the level of the institution rather than the individual) is the collection of material on workhouses that has been compiled by Peter Higginbotham and the Hospital Records database, sponsored jointly by the Wellcome Trust and TNA, which covers over 3,000 entries, providing administrative details of hospitals and the location of the archives associated with them.[139]

Vaccination registers and infant death registers have survived for some areas and a list of the known registers by sub-district is included in Drake and Razzell's *The decline of infant mortality in England and Wales 1871-1948*.[140] These registers can be extremely informative on matters relating to a mother's fertility history, as well as the medical and health development of new-born children. Some of the variables recorded apply at the household level as well as the individual level, such as information of accommodation and water supply.

## 7.3. Household level

Additional data at the household level would primarily relate to the property in which a sample household under observation were resident, such as ownership, value, or physical characteristics of the property in question. The biggest problem of linking information at the household level, especially in urban areas, is that address information may not remain constant over time, and in rural areas the address in the CEB is often too vague to pinpoint property. In the nineteenth-century not only did street names change, but house numbers were not always fixed.[141] As a result, even where household level sources exist, linkage to the VPS sample may not be without problems.

### *7.3.1. Rate books*

The main candidate for additional information at the household level is most probably local authority rate books, which were required (in various forms) following the Poor Law Amendment Act of 1834. These survive for many towns and cities and provide information on the value of the property in which

[139] See http:/www.workhouses.org.uk and http:/www.nationalarchives.gov.uk/hospitalrecords.
[140] Drake and Razzell, *The decline of infant mortality*.
[141] For a discussion on this important issue see Schürer and Mills, 'Residential patterns', 348-9.

the household lived.[142] Since rate books were often compiled on a more regular basis than census returns, in some instances on an annual basis, they could be useful for urban areas in providing information on the timing of intra-censual residential moves.[143] However, few seem to exist in computerised form, and most are housed in local archives and libraries, which would make a systematic linking exercise practically difficult and potentially costly.

### *7.3.2. Land tax assessments and manor court books*

Like rate books, in rural areas especially, land tax assessments and manor court books survive into the twentieth century for many rural areas, the latter detailing changes of property ownership.[144] Equally like rate books, these sources are mainly available only from local repositories, few appear to be indexed and even fewer are transcribed in machine-readable form.

### *7.3.3. Valuation returns*

Sometimes referred to as the New Domesday Survey, the 1910 Valuation Office survey was produced as a result of Lloyd George's 1909-10 Finance Act. Although outside of the period for the initially proposed VPS, it is included here for two reasons. First, should the VPS be funded, linking it through to 1911 once the census of that becomes available in 2012, is a logical development. Second, and moreover, the source is both national and incredibly rich in content. Details are included on property size, condition, ownership, value, tenure, occupancy and use.[145] Although records from this survey are available from TNA (with some copies also available locally), they suffer from two main drawbacks. Not only are they quite difficult to use, in terms of identifying individuals properties, but virtually none have been computerised. However, by 2012 this situation may well have changed, especially since TNA are keen (resources permitting) to digitise and index them.

### *7.3.4. Tithe returns*

Tithe apportionments were created as a result of the Tithe Commutation Act, 1836. These usually consist of a map showing properties in a given parish

[142] For an example of where rates books have effectively been linked to CEBs see Redfern, 'An early Victorian suburban élite'.
[143] Dennis, 'Intercensal mobility'.
[144] For the location of surviving land tax assessments see Gibson and Mills, *Land Tax Assessment*. However, see Stephens, *Sources for English local history,* 187-90 on the problems associated with using them.
[145] Short and Reed, *Landownership and society*; Short, *Land and Society in Edwardian Britain*; Short, *The Geography of England and Wales*.

liable to pay tithes and an accompanying schedule which details the newly agreed liabilities. Most apportionments created under the Act relate to the period c. 1840-1855.[146] Their importance from the point of view of the VPS is that the schedules list landowners and occupiers of properties, which could potentially be linked to the base 1851 sample, providing information on landownership and occupancy. However, coverage of the tithe returns is not complete, being concentrated in the southern and mid-England, and being essentially rural rather than urban in character. Equally, there is no complete computerised version of the information contained within the schedules.[147]

## 7.4. Area level

Of the three categories used here, area level data are seen, for obvious reasons, as the easiest to link to a VPS, although changes in administrative geographies over time would undoubtedly cause some difficulties. For the second half of the nineteenth-century a wealth of information exists at an area level across the whole country, yet some differences will inevitably occur between Scotland and its southern neighbours due to separate administrative responsibilities and reporting procedures. It is also the case that much nineteenth-century area data already exists in computerised form.

### *7.4.1. Aggregate census and civil registration statistics*

Whilst the fundamental building blocks of the VPS will be nominal level census and civil registration data, the GROs who were responsible for collecting these data, both north and south of the border, did so in effect to publish reports based on their analysis, much of which were presented in tabular form. To date, two major projects have worked to digitise the main outputs of this process, both of which have or are producing material that could be linked to VPS at an area level to provided contextual information on a range of socio-economic topics, such as: employment statistics, housing density, prevailing mortality and fertility levels, net migration.

#### 7.4.1.1. The Online Historical Population Reports (OHPR) project

The OHPR project is a JISC-funded digitisation project undertaken and recently completed by AHDS History at the UKDA. This created a digital version of the complete published population statistics for the United Kingdom for the period 1801 to 1931, scanning all the published reports arising from the

[146] Kain and Oliver, *The Tithe maps and apportionments;* Kain and Prince, *The Tithe Surveys*, and Evans, *The contentious tithe*. See also Mills, 'The residential propinquity of kin' and Henstock, 'House repopulation' as examples of where tithe apportionments have been linked to census enumerators' books.

[147] Yet see Kain, *Atlas of Agriculture*.

census and the annual and decennial reports of the Registrars-General of England and Wales, Scotland and Ireland, a total of approximately 190,000 images.[148] Whilst the main dissemination tool for the project is an on-line, web-based user interface that allows browsing, searching, viewing and downloading of the images, the project is also creating and making available machine-readable versions of a number of key statistical tables contained within the reports. It is these tables which could be linked to the VPS at the area level to provide contextual socio-economic information.

#### 7.4.1.2. The Great Britain Historical Database (GBHD)

GBHD is a large database of aggregate statistical information mainly covering the second half of the 19th and the early 20thcenturies, created by Dr Humphrey Southall and associates.[149] Whilst like the OHPR project it draws heavily on the published report of the census offices and Registrar Generals, it also contains tabulated statistics from the Poor Law Board as well as small debt statistics from county courts, 1847-1913. The GBHD is held by the UKDA and is in a form which could be linked to a VPS without significant difficulty.

### *7.4.2. Crime Statistics*

Crime statistics are published by area for crimes which came before the Assizes or Quarter Sessions annually since 1834 and for Judicial offences (crimes dealt with by magistrates) annually since 1857. These are classified under six headings:

- offences against the person e.g. murder, manslaughter, rape, assault;
- offences against property with violence e.g. burglary, robbery;
- offences against property without violence e.g. larceny, theft, embezzlement;
- malicious offences against property e.g. arson, machine breaking;
- forgery and offences against currency;
- other offences e.g. treason, smuggling, poaching, perjury, riot, sedition.[150]

TNA also hold a range of criminal statistics e.g. Statistics of Crime, 1881-92.[151] The Report of the Commissioners of Inland Revenue for 1869 gives a return of the number of houses licensed for the sale of intoxicating liquors.[152]

---

[148] See http://www.histpop.org/.

[149] The Great Britain Historical Database, compiled by H. Southall and colleagues, is available as a number of files from AHDS History at the UK Data Archive. See http://ahds.ac.uk/history/collections/census-statistics.htm.

[150] Hoyle, *Crime in England and Wales*.

[151] TNA HO 45/10424/R19175.

[152] TNA IR 15. This series begins in 1856.

While most of these statistics are available in Parliamentary Papers that have been scanned[153] it appears that very few area level series of criminal statistics are available in computerised form.

### *7.4.3. Education Statistics*

The 1851 Educational Census for England and Wales is available via the OHPR project (see section 7.4.1.1, above). Scanned images of the report and the detailed tables, giving number of schools and scholars by registration district, have been produced. TNA hold a range of educational statistics. The parish files 1872-1904 [ED 2] give the educational census returns from outside London, while the London Educational Returns [ED 3] give types of school and pupil numbers for London. The minutes and reports of the committee of the Privy Council 1839-99 [ED 17] detail the provision of elementary education and attendance. There are educational census returns in the Public Elementary School Files [ED 21].[154] Whilst potentially informative, none of these sources as yet are available in a database format.

### *7.4.4. Religious statistics*

The report for the 1851 Census of Religious Worship is available from the OHPR project (see above). There are scanned images of the report and the detailed tables, which give numbers of places of worship, numbers of sittings and numbers of attendees by denomination by registration district. Abstracted statistics are also available from the GHDB (see above). Snell and Ell have analysed these data and produced a series of about 40 maps and a database covering part, but not all of the country.[155] There were no further national religious censuses but the *Nonconformist* organized and published the results of censuses of London in 1865 and large towns [population greater than 20,000] in 1872, giving details of Anglican and Roman Catholic as well as Protestant non-conformist places of worship.[156] The results of a private census carried out by newspapers in about 125 towns were published in 1882.[157]

### *7.4.5. Meteorology statistics*

The National Meteorological Library and Archive began in the 1850s as the library of the Meteorological department of the Board of Trade and provides a

[153] See http://www.bopcris.ac.uk.
[154] See also Morton, *Education and the state from 1833*.
[155] Snell and Ell, *Rival Jerusalem*.
[156] Stephens, *Sources for English Local History*, p.276; *The Nonconformist*, 15 November, 1865 and 23 October, 1872.
[157] Mearns, *The statistics of attendance at public worship*.

record of past weather and climate conditions.[158] Their observations from a thousand sites date back to the mid-nineteenth century, covering temperature, wind, rainfall, solar radiation, snow and sunshine. Although one would expect these to be computerised, no record of this could be found.

### *7.4.6. Medical Officer of Health Reports*

Reports made by the various Medical Officers of Health, where they have survived, provide valuable area level information and may also include information at the household level. In addition to information on housing density and mortality rates, reports can provide details on morbidity, sewerage disposal, drainage and water supply. Between 1848 and 1857 local reports to the General Board of Health were published for about 300 places. The most comprehensive collection is believed to be at the British Library, although the Wellcome Institute Library also holds a significant collection.[159]

Although these reports contain some fascinating information not available from other sources, the ability to link these to a VPS would be troublesome, not only because there coverage is partial and varies over time, but also because the information recorded within them is not reported systematically and virtually nothing is available in computerised form.

### *7.4.7. Newspapers*

Newspapers, while providing a lot of area level information and some individual level data, are a difficult source to use. They are widely available locally but are rarely indexed. The London Gazette can only be searched online for the twentieth century.[160] One example of where newspaper sources have been used is that co-ordinated by FACHRS which has successfully run a project extracting information about incidents associated with the Swing disturbances. This project produced information from over a dozen counties, covering a three-year period, in a very short time, showing what can be done by volunteers working under the direction of a project co-coordinator.[161]

## 7.5. Other sources

In addition to the various sources mentioned so far, a number of sources could be seen as possible candidates for linking to a VPS. These include:

- pension records;
- army lists;

[158] See http://www.met-office.gov.uk.
[159] Stephens, *Sources for English Local History*, p.320
[160] See http:/www.gazettes-online.co.uk.
[161] See hhtp://www.fachrs.com/swing/swing_projects.htm.

- school registers;
- Friendly Society membership data;
- crews of incoming and outgoing ships;
- emigration records;
- foreign CEBs.

Old age pension records date from 1908, after the period covered by VPS, although if a VPS were to be extended to 1911, these might become relevant. Military pensions were paid throughout the period covered by VPS. The records of these are at TNA in classes PMG12 and WO24 or 25. Civil service pensions were paid to a limited number of people. Records of some of these are held in the British Library India Office records, the House of Lords record office, and in the Royal Mail archives.

Army lists are available at TNA in class WO32. Service registers of Royal Naval Seamen between 1853 and 1923 are now available online from TNA.

Some school registers have been deposited in county record offices but their survival is patchy. Only 13, mainly for the Royal Greenwich Hospital School, are deposited at TNA.

There is some material on Friendly Societies at TNA class FS 15. More detail can probably be found among quarter sessions records at county record offices and at the Modern Records Centre, Warwick University.

There are no comprehensive immigration or emigration records for the period. Such records as exist are among the Colonial Office papers at TNA and among the India Office Records at the British Library. Agreements and crew lists are available at TNA among Board of Trade papers, class BT98.

None of the sources discussed above are systematically available in computerised form, and are either of patchy coverage or limited to certain sub-groups of the population. Thus they would all be of a low priority. Of the sources mentioned in the User Survey, linkage between the VPS and censuses of other countries offers the greatest possibilities. There are five countries in the world that possess completely digitised individual-level censuses for the late nineteenth century: Canada, Great Britain, Iceland, Norway, and the United States. A project already exists to cross harmonise these censuses through the creation of common coding schemes (see section 6.4), with the University of Essex as the UK project partner.[162] Work has already started at the Minnesota Population Center on examining the possibility of linking individuals across these census data and the VPS project cold benefit from this initiative.

## 7.6. Conclusions

Although the notion of enhancing a VPS by linking additional sources at either the individual, household or area levels is highly attractive, the initial priority

[162] The North Atlantic Population Project (NAPP). See http://www.nappdata.org/.

must be with creating the VPS itself. Thus any linking of additional sources must be seen as a secondary goal at best. Of the possible candidates linking pre-existing area level data from the GBHD and OHPR projects is seen as both practical and desirable in order to provide important contextual information. Beyond this, at the individual level, back-linking to parish register data may prove beneficial in order to complete fertility and marital histories of sample members prior to 1851.

Linkage of most of the other sources mentioned above, even where they exist in computerised form, would probably be relatively time consuming and thus costly at the national level. However, one possibility that exists would be to generate a small number of enriched 'community' sub-samples. This could perhaps be funded separately on a case-by-case basis and could well benefit from the input of volunteers from among the local and family history groups in order to offset costs by the use of volunteers communities.

## 8. Recommendations and Future Strategy

If a VPS is to be created then it is strongly recommended that the creation of it proceed as a staged process. This is suggested since the landscape in which a VPS project would operate will undoubtedly change in the coming years, and the project as a whole should be flexible in order to adapt to these changes. A staged approach also makes sound financial sense. Each stage, as suggested below, would have distinct and independent advantages to the research community. Thus, funding stage one, does not necessary require a commitment to fund either stages two or three. Instead progress through the various stages can be evaluated independently and the cumulative investment assessed at each stage before proceeding to the next.

Drawing on the findings reported in this publication, what follows is a stage-by-stage recommendation of how a VPS might best be created.

Stage 1: *Create fully coded census micro data*

As has been detailed already, it will be impossible to create a linked VPS database unless the underlying data have been fully cleaned, standardised and coded. This is a fundamental requirement in any detailed record linkage exercise. Thus, as a first stage it is recommended that fully coded and harmonised versions of the 100 per cent census micro data for England and Wales for 1851, 1861, 1871, 1881, 1891 and 1901, in total a data resource covering some 148 million person records, are created. Regardless of any linking work done thereafter, this will provide a tremendous research resource in its own right, essentially generating a GB version of the US-based IPUMS project, albeit with a different time span and including full rather than sample data.

Work on stage one could start almost immediately. This initial project would build on work already undertaken on the 1861 and 1881 censuses, and indeed could not be achieved within such a short time frame if it where not able to build on the existing foundations that the work on these two census years has established already. The 1891 are immediately available from findmypast.com (formerly 1837online.com) and they expect that the data for 1871 will be available during the course of 2007 and those for 1851 during 2008. The data for 1901 are also currently available, pending an agreement with FriendsReunited.
Work on Stage 1 would essentially consist of six key elements:

- *Workpackage 1* – Re-formatting, checking and cleaning the existing data
- *Workpackage 2* – Preparing a harmonised multi-level occupation coding schema, mapped to NAPP, ISCO88, SCO2000, NS-SEC and other relevant classifications, including the Registrars-Generals classifications of 1851 to 1911.
- *Workpackage 3* – Preparing a harmonised census enumeration geography for 1851 to 1901.
- *Workpackage 4* – Coding and standardising all fields recorded in the census data.
- *Workpackage 5* – Creation of a parish-level GIS to link seamlessly to the census data.
- *Workpackage 6* – Creation of representative national samples, documentation, user guides, supporting teaching and research materials.

Stage 2: *Code indexed data*

Once the main census data have been fully coded and harmonized, attention can be turned to the civil registration data and the indexed census data for Scotland. These will also need to be re-formatted, checked, cleaned and coded. There is no imperative for these to be worked on at the same time as the bulk of the English and Welsh census data. Indeed, there are clear advantages in waiting, first to enable FreeBMD (a charity who are transcribing the entire civil register index entries) and ONS to complete there work on the indexing of the civil registers for the period, and second, there is a possibility that over time GRO(S) may enrich some or all of their indexed census data. It is estimated that the task of clearing and coding these data could take around a year. As with Stage I, once standardised, both sets of data will provide a valuable research resource in their own right.
Work on Stage 2 would essentially consist of five key elements:

- *Workpackage 1* – Re-formatting, checking and cleaning the existing data.
- *Workpackage 2* – Coding and standardising all fields recorded in the census data.
- *Workpackage 3* – Standardising entries in the civil registration data (especially geography and names).
- *Workpackage 4* – Creation of 'dummy' marriages from the civil registration data.

- *Workpackage 5* – Producing necessary documentation, user guides, support materials.

Stage 3: *Work on preparing 1911 census data*

Although the original intention was to create a VPS covering the period 1851 to 1901, given the plans to release the 1911 census in 2012, a logical step would be to extend the VPS to 1911. Whilst this could be done as a completely separate exercise after an initial 1851 to 1901 VPS is created, if the timing is appropriate, it would be cost efficient to prepare these data before linking starts and incorporate them into the main linking exercise. Adding another census year into the main linking process could be achieved at virtually no additional cost. However, to link 1851 to 1901 and then link in 1911 separately at a later date would clearly add to total costs.

Clearly, creating a useable version of the 1911 census would provide a significant research resource, irrespective of any further linking, especially given the significance of the 1911 census for studies into the fertility and mortality transitions.[163]

Work on Stage 3 would essentially consist of three key elements, assuming that enumeration geography standardisation and GIS work for 1911 were incorporated into Stage 1:

- *Workpackage 1* – Re-formatting, checking and cleaning the existing data
- *Workpackage 2* – Coding and standardising all fields recorded in the census data.
- *Workpackage 3* – Creation of representative national samples, documentation, user guides, supporting teaching and research materials.

Stage 4: *Link data*

Once all the underlying data have been checked, cleaned, re-formatted, standardised and coded, then the VPS project could move onto the main linking exercise, and thereby create a longitudinal database of huge and important research potential.

Work on Stage IV would follow the linking strategy set out in section 7 above, and it seems reasonable to allocate a minimum of two years to complete these tasks.

Stage 5: *Enhance linked data*

Given that under the present situation liking will have to be carried out on indexed civil registration data, and partial transcriptions of some of the Scottish census data, although it is not entirely necessary, it would be highly desirable to enhance the linked data by adding in details not included on the indexes and partial transcriptions. This will greatly add to the value of the longitudinal dataset, providing rich information on VPS members (such as cause of death, occupation and address at time of marriage) that would otherwise be lost. In addition, area level information, used to contextualise the longitudinal data should also be added in at stage. It also assumes that the full certificate-level historic civil registration information for

[163] Garrett et al, *Changing family size*.

England and Wales will be made freely-available to the VPS on the same terms as has been suggested by GRO(S).

Work on Stage 5 would essentially consist of five key elements, as follows:

- *Workpackage 1* – Locating VPS members in the civil registration material and transcribing additional information.
- *Workpackage 2* – Locating VPS members in the partially transcribed Scottish census information and transcribing additional information.
- *Workpackage 3* – Coding and standardising all newly entered field.
- *Workpackage 4* – Creation of area level variables and joining them to VPS.
- *Workpackage 5* – Re-formating of VPS database, creation of documentation, user guides, supporting materials.

# 9. Bibliography

## 9.1. Data Sources


Anderson, M. *et al*. *National Sample from the 1851 Census of Great Britain* [computer file]. Colchester, Essex: UK Data Archive [distributor], 1979. SN: 1316.

City University. Social Statistics Research Unit, *National Child Development Study Composite File Including Selected Perinatal Data and Sweeps One to Five, 1958-1991* [computer file]. *2nd Edition.* National Birthday Trust Fund, National Children's Bureau, City University. Social Statistics Research Unit, [original data producer(s)]. Colchester, Essex: UK Data Archive [distributor], 2000. SN: 3148.

Elliott, J., *1970 British Cohort Study: Partnership Histories, 1986-2000* [computer file]. Colchester, Essex: UK Data Archive [distributor], 2005. SN: 5218.

French, C., Sullivan, A. and Warren, J., *Burial Registers for Kingston upon Thames Parishes, 1850-1901 and Bonner Hill Cemetery, 1855-1911* [computer file]. Colchester, Essex: UK Data Archive [distributor], 2002. SN: 4423.

Joint Centre for Longitudinal Research, *National Child Development Study and 1970 British Cohort Study (BCS70) Follow-ups, 1999-2000* [computer file]. *2nd Edition.* Colchester, Essex: UK Data Archive [distributor], 2003. SN: 4396.

Kain, R.J.P., *Atlas of Agriculture in England and Wales, c.1840* [computer file]. Colchester, Essex: UK Data Archive [distributor], 1981. SN: 1659.

Kirkman, K., *Pinner Census, 1841-1891* [computer file]. Colchester, Essex: UK Data Archive [distributor], 1993. SN: 2988.

McCormick, M., *Cornwall Online Census Project, 1891* [computer file]. Colchester, Essex: UK Data Archive [distributor], 2004. SN: 4978.

Marmot, M. *et al*., *English Longitudinal Study of Ageing: Wave 1, 2002-2003* [computer file]. *3rd Edition.* Colchester, Essex: UK Data Archive [distributor], 2005. SN: 5050.

Rau, D., *1891 Census Project, Spitalfields* [computer file]. Colchester, Essex: UK Data Archive [distributor], 1994. SN: 3139.

Schürer, K. and Woollard, M. *1881 Census for England and Wales, the Channel Islands and the Isle of Man (Enriched Version)* [computer file]. Genealogical So-

ciety of Utah, Federation of Family History Societies [original data producers]. Colchester, Essex: UK Data Archive [distributor], 2000. SN: 4177.
Schürer, K. and Woollard, M. *1881 Census for Scotland (Enriched Version)* [computer file]. Genealogical Society of Utah, Federation of Family History Societies [original data producers]. Colchester, Essex: UK Data Archive [distributor], 2000. SN: 4178.
Southall, H.R., Gilbert, D.R. and Gregory, I. *Great Britain Historical Database Online, 1841-1939* [computer file]. Colchester, Essex: AHDS History, UK Data Archive [distributor], 2000. SN: 33305.
Tilley, P. et al., *Census Enumerators' Returns for Kingston upon Thames, 1851, 1861, 1871 and 1891* [computer file]. Colchester, Essex: UK Data Archive [distributor], 2004. SN: 4710.
University of Essex. Institute for Social and Economic Research, *British Household Panel Survey; Waves 1-13, 1991-2004* [computer file]. Colchester, Essex: UK Data Archive [distributor], 2005. SN: 5151.
University of London. Institute of Education. Centre for Longitudinal Studies, *Millennium Cohort Study: First Survey, 2001-2003* [computer file]. *3rd Edition.* Colchester, Essex: UK Data Archive [distributor], 2004. SN: 4683.

## 9.2. Publications referenced in text of the report

Alter, G. *Family and the female life course: the women of Verviers, Belgium, 1849-1880*, (Madison, 1988).
Anderson, M. 'Marriage patterns in Victorian Britain: an analysis based on registration district data for England and Wales, 1861', *Journal of Family History*, 1 (1976) 55-79.
Anderson, M., 'Standard tabulation procedures for the census enumerators' books, 1851-1891', in Wrigley, ed, *Nineteenth-century society*, 134-45.
Armstrong, W. A., 'The use of information about occupation', in Wrigley, ed, *Nineteenth-century society*, 191-310.
Baines, D. *Migration in a mature economy: emigration and internal migration in England and Wales, 1861-1900*, (Cambridge, 1985).
Bouchard, G., 'The processing of ambiguous links in computerized family reconstruction', *Historical Methods,* 19 (1986) 9-19.
Burton, V. C. 'A floating population: vessel enumeration returns in censuses, 1851-1921' in Mills and Schürer, eds., *Local communities,* 47-55.
Cheffins, R.H.A., *Parliamentary constituencies and their registers since 1832: a list of constituencies from the Great Reform Act with the British Library's holdings of electoral registers together with the Library's holdings of burgess rolls, poll books and other registers*, (London, 1998).
Copas, J. and Hilton, F., 'Record linkage: statistical models for matching computer records', *Journal of the Royal Statistical Society*, 153 (1990) 287-320.
Crompton, C. A. 'An Exploration of the Craft and Trade Structure of Two Hertfordshire Villages, 1851-1891: an Application of Nominal Record Linkage to Directories and Census Enumerators Books', *The Local Historian,* 28 (1998) 145-58.
Davies, R., Garrett, E. and Reid, A., 'Nineteenth century Scottish demography from linked census and civil registers', paper to be presented at the International Mi-

crodata Access Group workshop (IMAG) – Longitudinal and Cross-sectional Historical Data: intersections and opportunities, Montreal, 10/11 November 2003.

Dennis, R. J. 'Intercensal mobility in a Victorian city', *Transactions of the Institute of British Geographers*, new series **2** (1977) 349-63.

Drake, M. and Razzell, P. *The decline of infant mortality in England and Wales 1871-1948: a medical conundrum*, (Milton Keynes, 1997).

Elfeky, M. G., Verykios, V. S. and Elmagarmid, A. K., 'Tailor: A record linkage toolbox' in the Proceedings of the 18th International Conference on Data Engineering (ICDE 2002), San Jose, California, 2002. Available at: http://citeseer.ist.psu.edu/article/elfeky02tailor.html.

ESRC. *Strategic Plan 2005-2010*, (Swindon, 2005).

Evan, E. J., *The Contentious Tithe*, (London, 1976).

Ferrie, J., 'A new sample of males linked from the Public-Use-Microdata-Sample of the 1850 US Federal Census of Population to the 1860 US Federal Census Manuscript Schedules', *Historical Methods*, 29 (1996) 141-56.

Fortini, M., Liseo, B., Nuccitelli, A., Scanu, M., 'On Bayesian record linkage', *Sixth International World Meeting on Bayesian Analysis,* (2000).

Friedlander, D. and Roshier, R. J. 'A Study of Internal Migration in England and Wales: Part 1', *Population Studies,* 19 (1966) 239-79.

Fure, E., 'Interactive record linkage: the cumulative construction of life courses', *Demographic Research,* 3 (2000). Available at: www.demographic-research.org.

Garrett, E. 'Trials of labour: motherhood versus employment in a nineteenth-century textile centre', *Continuity and Change*, 5 (1990) 121-54.

Garrett, E., Reid, A., Schürer, K. and Szreter, S., *Changing family size in England and Wales. Place, class and demography, 1891-1911*, (Cambridge, 2001).

Gibson, J. and Rogers, C., *Poll books c.1696-1872: a directory to holdings in Great Britain*. 3rd ed., Federation of Family History Societies, (Birmingham,1994).

Gibson, J. and Mills, D. *Land Tax Assessment c.1690-c.1950,* Federation of Family History Societies, (Birmingham, 1983).

Gibson, J. and Rogers, C., *Electoral registers since 1832*. 2nd ed. Federation of Family History Societies, (Birmingham, 1990).

Goldthorpe, J.H., 'The 'Goldthorpe' class schema: some observations on conceptual and operational issues in relation to the ESRC review of government social classifications' in Rose and O'Reilly, eds, *Constructing classes.*

Goldthorpe, J.H., *Social mobility and class structure in modern Britain*, (Oxford, 1980).

Goose, N., 'Workhouse populations in the mid-nineteenth century: the case of Hertfordshire', *Local Population Studies*, 62 (1999) 52-69.

Guth, G. J. A. 'Surname spelling and computerized record linkage', *Historical Methods Newsletter,* 10 (1976) 10-9.

Hancock, R. 'In service or one of the family? Kin-servants in Swavesey 1851-1881, Ryde 1881, and Stourbridge 1881', *Family and Community History,* 2 (1999) 141-18.

Harvey, C. and Green, E., 'Record linkage algorithms: efficiency, selection and relative confidence', *History and Computing,* 6 (1994) 143-52.

Harvey, C., Green, E. and Corfield, P.J., 'Record linkage theory and practice: an experiment in the application of multiple pass linkage algorithms' *History and Computing,* 8 (1996) 78-89.
Hattersley, L. and Creeser, R. *Longitudinal study 1971-1991. History, organisation and quality of data*, OPCS Series LS no. 7 (London, 1995).
Henstock, A., 'House repopulation from the CEBs of 1841 and 1851', in Mills and Schürer eds, *Local communities*, 363-82.
Higgs, E. *A clearer sense of the census. The Victorian censuses and historical research*, (London, 1996).
Higgs, E., 'Structuring the past: the occupational, social and household classification of census data', *Computing and History Today*, 4 (1988) 24-30.
Hinde, A. and Turnbull, F., 'The population of two Hampshire workhouses, 1851-61', *Local Population Studies*, 61, (1998) 38-53.
Hinde, P. R. A. 'The population of a Wiltshire village in the nineteenth century: a reconstruction study of Berwick St James, 1841-1871', *Annals of Human Biology,* 14 (1987) 475-85.
Hoyle, W. *Crime in England and Wales: an historical and critical retrospect*, (London, 1876).
Hurren, E., 'Welfare-to-work schemes and a crusade against outdoor relief in the Brixworth Union, Northamptonshire, in the 1880s', *Family and Community History*, 4 (2001) 19-30.
International Labour Office, *International standard classification of occupation*. Revised edition 1968, (Geneva, 1969).
International Labour Organisation, *International classification of occupation, 1988*, (Geneva, 1990).
Jackson, D.G., 'Kent workhouse populations in 1881: a study based on the CEBs', *Local Population Studies*, 69 (2002) 51-66.
Jackson, D.G., 'The Medway Union Workhouse, 1876-81: a study based on the admission and discharge registers and the CEBs', *Local Population Studies*, 75 (2005) 11-32.
Jaro, M. A., 'Probabilistic linkage of large public health datafiles', *Statistics in Medicine*, 14 (1995) 491-8.
Kain, R. J. P. and Oliver, R. R., *The Tithe maps and apportionments of England and Wales*, (Cambridge, 1994).
Kain, R. J. P. and Prince, H. C., *The Tithe surveys of England and Wales*, (Cambridge, 1985).
Katz, M. and Tiller, J. 'Record linkage for everyman: a semi-automatic process', *Historical Methods Newsletter,* 5 (1972) 144-50.
King, S., 'Multiple-source record linkage in a rural industrial community, 1680-1820', *History and Computing,* 6 (1994) 133-42.
King, S., 'Record linkage in a proto-industrial community', *History and Computing,* 4 (1992) 27-33.
Lait, A. and Randell, B., 'An assessment of name matching algorithms', Department Technical Report, series no. 550, Department of Computing Science, University of Newcastle upon Tyne.
Larsen, M. D. and Rubin, D. B., 'Iterative automated record linkage using mixture models', *Journal of the American Statistical Association*, 96 (2001) 32-41.

Lawton, R. 'Population changes in England and Wales in the later nineteenth century: an analysis of trends by registration districts', *Transactions of the Institute of British Geographers*, 44 (1968) 55-74.
Long, J., 'Urbanisation, internal migration, and occupational mobility in Victorian Britain'. [unpublished typescript, November 2001, provided to author.]
Mackinon, M., 'English Poor Law policy and the crusade against outdoor relief', *Journal of Economic History*, 47 (1987) 603-25.
Mandemakers, K. and Boonstra, O., *De levensloop van de Utrechtse bevolking* (Assen 1995).
Mearns, A., *The statistics of attendance at public worship*, (London, 1882).
Mills, D. R. and Pearce, C. *People and places in the Victorian census. A review and bibliography of publications based substantially on the manuscript census enumerators' books, 1841-1911*, Institute of British Geographers, Historical Geography Research Series, No 23, (Cheltenham, 1989).
Mills, D. R. and Schürer, K. eds *Local communities in the Victorian census enumerators' books*, (Oxford, 1996).
Mills, D. R. and Schürer, K., 'Employment and occupations' in Mills and Schürer, eds, *Local communities*, 136-60.
Mills, D. R., 'The residential propinquity of kin in a Cambridgeshire village, 1841', *Journal of Historical Geography*, 4 (1978), 265-76.
Mineau, G. P., Bean, L. L. and Anderton, D. L. 'Description and evaluation of linkage of the 1880 census to family genealogies: implications for Utah fertility research', *Historical Methods,* 22 (1989) 144-57.
Morton, A., *Education and the state from 1833*, (Kew, 1997).
Nenadic, S. 'Studying the middle class in nineteenth-century urban Britain', *Urban History Yearbook*, (1987) 22-41.
Nygaard, L., 'Name standardization in record linking: an improved algorithmic strategy', *History and Computing,* 4 (1992) 63-74.
Office for National Statistics, *Civil registration: delivering vital change*, (London, 2003).
Office for National Statistics, *Standard occupational classification 2000*. Volume 1: structure and description of unit groups, (HMSO, London, 2000).
Office for National Statistics, *Standard occupational classification 2000*. Volume 2: the coding index, (HMSO, London, 2000).
Olsson, F. and Reuterswärd, E., *Skånes demografiska databas 1646-1894. En källbeskrivning*. Lund Papers in Economic History, 32 (1993).
O'Reilly, K., and Rose, D., eds, *Constructing classes: towards a new social classification for the UK*, (London, 1997).
Perkyns, A. 'Age checkability and accuracy in the censuses of six. Kentish parishes, 1851-1881', in Mills and Schürer eds *Local communities*, 115-34.
Perkyns, A. 'Birthplace accuracy in the censuses of six. Kentish parishes, 1851-1881', in Mills and Schürer eds *Local communities*, 229-45.
Pevalin, D. J. and Rose, D., eds, *A researcher's guide to the National Statistics Socio-economic Classification*, (London, 2003).
Pooley, C. G. 'Residential mobility in the Victorian city', *Transactions of the Institute of British Geographers*, new series 4 (1979) 258-77.

Pooley, C. G. and Doherty, J. C. 'The longitudinal study of migration. Welsh migration to English towns in the nineteenth century', in Pooley and Whyte eds *Migrants, emigrants and immigrants*, 143-73.
Pooley, C. G. and Turnball, J. Migration and mobility in Britain since the eighteenth century, (London, 1998).
Pouyez, C., Roy, R. and Martin, F., 'The linkage of census name data: problems and procedures', *Journal of Interdisciplinary History*, 14 (1983) 129-52.
Reay, B. 'Kinship and the neighbourhood in nineteenth-century rural England: the myth of the autonomous nuclear family', *Journal of Family History,* 21 (1996) 87-104.
Redfern, J. B., 'An early Victorian suburban elite: heads of household at home', in Mills and Schürer, *Local communities*, 394-407.
Redmonds, G., *Christian names in local and family history*, (Kew, 2004).
Roberts, D. and Roberts, M., 'Surnames and relationships: an Orkney study', *Human Biology* 55 (1983) 341-7.
Roberts, E., Dillon, L. Y., Woollard, M., Thorvaldsen, G., and Ronnander, C., 'Occupational classification in the North Atlantic Population project', *Historical Methods*, 36 (2003) 89-96.
Robertson, A. M. and Willett, P., 'Searching for historical word-forms in a database of 17th century English text using spelling-correction methods', (1992). Available at: http://widit.slis.indiana.edu/irpub/SIGIR/1992/pdf24.pdf
Rogers, C., *The surname detective: investigating surname distribution in England 1086 – present day*, (Manchester, 1995).
Rose, D., Pevalin, D. J. and O'Reilly, K., *The NS-SEC: origins, development and use*, (London, 2005).
Ruggles, S., 'Linking historical censuses: a new approach', IMAG workshop, Montreal, Nov. 2003.
Schofield, R. S., 'Automatic family reconstitution: the Cambridge experience', *Historical Methods,* 25 (1992) 75-9.
Schofield, R. S., 'The standardization of names and the automatic linking of historical records', *Annales de Démographie Historique,* (1972) 359-64.
Schürer, K., 'Historical demography, social structure and the computer', in Denley, P. and Hopkin, D., eds, *History and Computing*, (Manchester, 1987) 33-45.
Schürer, K. 'The role of the family in the process of migration', in Pooley, C. G. and Whyte, I. D., eds, *Migrants, emigrants and immigrants*. *A social history of migration*, (London, 1991) 106-42.
Schürer, K., 'Understanding and coding the occupations of the past', in Schürer, K. and Diederiks, H., eds, *The use of occupations in historical analysis*, (St. Katharinen, 1993) 101-62.
Schürer, K., 'The Victorian Panel Survey: a scoping study for the ESRC', (unpublished report, October 2003).
Schürer, K. and Dillon, L., 'What's in a name? Victorias in late nineteenth-century Great Britain and Canada', *Local Population Studies*, 70 (2003) 57-62.
Schürer, K. and Mills, D. R., 'Family and household structure', in Mills and Schürer, eds, *Local communities*, 280-97.
Schürer, K. and Mills, D. R., 'Residential patterns', in Mills and Schürer, eds, *Local communities*, 348-62.

Schürer, K., Oeppen, J. and Schofield, R. S., 'Theory and methodology: an example from historical demography', in Denley, P., Fogelvik, S. and Harvey, C., eds, *History and Computing II*, (Manchester, 1989) 130-42.
Short, B. and Reed, M., *Landownership and society in Edwardian England and Wales: the Finance (1909-10) Act 1910 records*, (Brighton, 1987).
Short, B., *Land and Society in Edwardian Britain*, (Cambridge, 1997).
Short, B., *The geography of England and Wales in 1910: an evaluation of Lloyd George's 'Domesday of Landownership'*, Historical Geography Research Series, no 22, (Cheltenham, 1989).
Sims, J. ed., *A handlist of British parliamentary poll books*, Occasional publication, University of Leicester History Department, no. 4, (Leicester, 1984).
Snae, C. and Diaz, B, 'An interface for mining genealogical nominal data using the concept of linkage and a hybrid name matching algorithm', (unpublished paper, Department of Computer Science, University of Liverpool, n.d.) Available at: http://www.csc.liv.ac.uk/~chakkrit/Publications/hc2001_Journal.pdf.
Snell, K. and Ell, P., *Rival Jerusalems* (Cambridge, 2000).
Society of Genealogists, *Directories and poll books including almanacs and electoral rolls in the Library of the Society of Genealogists*, 6th ed, (London, 1995).
Stephens, W. B., *Sources for English Local History* (Cambridge, 1994).
Szreter, S. R. S., 'The genesis of the Registrar-General's social classification of occupations', *British Journal of Sociology*, 35 (1984) 522-46.
Taylor, M. F. ed. with Brice, J., Buck, N. and Prentice-Lane, E., *British Household Panel Study user manual*. Volume A. Introduction, technical report and appendices, (Colchester, 2002).
Tetielbaum, M. S. *The British fertility decline*, (Princeton, 1984).
Thomson, D., 'Workhouse to nursing home: residential care of elderly people in England since 1840', *Ageing and Society*, 3 (1985), 43-69.
Tilley, P., 'The Kingston local history project: creating life histories and family trees for communities in Victorian Britain', International Microdata Access Group workshop (IMAG) – Longitudinal and Cross-sectional Historical Data: intersections and opportunities, Montreal, 10/11 November 2003.
Van Leeuwen, M. H. D., Maas, I., and Miles, A., *HISCO: Historical International Standard Classification of Occupations*, (Leuven, 2002).
Wall, W. D. and Williams, H. L. *Longitudinal studies and the social sciences*, Social Science Research Council. Reviews of Current Research, no. 7. (London, 1970).
Webb, C. and Hemingway, V., 'Improving access: a proposal to create a database for the probate records at the Borthwick Institute', *History and Computing,* **7** (1995) 152-5.
White, M. B., 'Family migration in Victorian Britain: the case of Grantham and Scunthorpe', in Mills and Schürer eds, *Local communities*, 267-77.
Wojciechowska, B., 'Brenchley: a study of migratory movements in a mid-nineteenth-century rural parish', in Mills and Schürer eds, *Local communities*, 253-66.
Woods, R. I., 'Approaches to the fertility transition in Victorian England', *Population Studies*, 41 (1987) 283-311.
Woods, R. I., *The demography of Victorian England and Wales*, (Cambridge, 2000).

Woollard, M., 'Conceptions of work: occupational classification in the British Isles, 1660-1911', (unpublished Ph.D, University of Essex, 2005).
Wrigley, E. A. and Schofield, R. S., *The population history of England 1541-1871: a reconstruction*, (Cambridge, 1981).
Wrigley, E. A., Davies, R. S., Oeppen, J. O. and Schofield, R. S., *English population history from family reconstitution, 1580-1837,* (Cambridge, 1997).
Wrigley, E. A., ed, *Nineteenth-century society: essays in the use of quantitative methods for the study of social history*, (Cambridge, 1972).
Zobel, J. and Dart, P., 'Finding approximate matches in large lexicons', *Software – Practice and Experience*, 25 (1995) 331-45.

## 9.3. Publications on record linkage consulted in reference to section 5 but not directly cited in text

Acheson, E. D., *Medical record linkage,* (Oxford, 1967).
Adman, P., Baskerville, S. W. and Beedham, K. F. 'Computer-assisted record linkage: or how best to optimize links without generating errors' *History & Computing,* 4 (1992) 2-15.
Alter, G., Gutmann, M., and Mandemakers, K., 'Problems and possibilities for distributing longitudinal historical data', Workshop on 'Disseminating and analyzing longitudinal historical data', International Institute of Social History, Amsterdam, March 2006.
Anderson, M. *National sample from the 1851 census of Great Britain: introductory user guide*, (University of Edinburgh, Dept. of Economic and Social History, 1987)
Anderson. M., Kemmer, D. and Morse, D. J., *Demographic change in Scotland 1855-1914: the onset of the fertility decline in Scotland, some results from an exercise in family reconstruction*, Working paper 2, (version 2), Department of Economic and Social History and University of Edinburgh Data Library, (University of Edinburgh, 1992).
Annal, D., *Using birth, marriage and death records*, (Richmond, 2002).
Atack, J., Bateman, F. and Gregson, M. E., '"Matchmaker, matchmaker, make me a match": a general personal computer-based matching program for historical research', *Historical Methods,* 25 (1992) 53-66. Available at: http://faculty.econ.northwestern.edu/faculty/ferrie/papers/Exceptionalism.pdf.
Beckett, J.V. and Foulds, T., 'Beyond the micro: Laxton, the computer and social change over time', *Local Historian* 15 (1985) 451-6.
Bengtsson, T. and Lundh, C., *Evaluation of a computer program for automatic family reconstruction*, Working paper presented at a seminar of the Cambridge Group for the History of Population and Social Structure, 6 May 1991.
Bengtsson, T. and Lundh, C., *Name-standardisation and automatic family reconstitution*. Lund Papers in Economic History, 29 (1993).
Bloothooft, G., 'Assessment of systems for nominal retrieval and historical record linkage', *Computers and the Humanities,* 32 (1998) 39-56.
Bloothooft, G., 'Corpus-based name standardization', *History and Computing,* 6 (1994) 153-67.

Bloothooft, G., 'Multi-source family reconstitution', *History and Computing,* 7 (1995) 90-103.
Bouchard, G., 'Current issues and new prospects for computerized record linkage in the Province of Quebec', *Historical Methods,* 25 (1992) 67-73.
Bynner, J., Butler, N., Ferrie, E., Shepherd, P. and Smith, K., *The design and conduct of the 1999-2000 survey of the National Child Development Study and the 1970 British Cohort Study*, Centre for Longitudinal Studies, Working Paper no. 1 (London, n.d.).
Condran, G. A. and Seaman, J., 'Linkage of the 1880-81 Philadelphia death register to the 1880 manuscript census: a comparison of hand and machine record linkage techniques', *Historical Methods,* 14 (1981) 73-84.
Crawford, E. M., *Counting the people. A survey of the Irish censuses, 1813-1911*, (Dublin, 2003).
Crompton, C. A., 'An exploration of the craft and trade structure of two Hertfordshire villages, 1851-1891: an application of nominal record linkage to directories and census enumerators books', *The Local Historian,* 28 (1998) 145-58.
Davies, H. R., 'Automated record linkage of census enumerators' books and registration data: obstacles, challenges and solutions', *History and Computing,* 4 (1992) 16-26.
De Brou, D. and Olsen, M., 'The Guth algorithm and the nominal record linkage of multi-ethnic populations', *Historical Methods,* 19 (1986) 20-4.
Dennis, R. J., 'Distance and social interaction in a Victorian city', *Journal of Historical Geography*, 3 (1977) 237-50.
Despotidon, S. and Shepherd, P., *1970 British Cohort Study twenty-six year follow-up. Guide to data available at the ESRC Data Archive*, (London, n.d.).
Dillon, L. and Desjardins, B., 'The historical demography research infrastructure: challenges and opportunities for census linkage in the French and Anglo Canadian context', International Microdata Access Group workshop (IMAG) – Longitudinal and Cross-sectional Historical Data: intersections and opportunities, Montreal, 10/11 November 2003.
Dillon, L. Y., 'International partners, local volunteers and lots of data: the 1881 Canadian census project', *History and Computing*, 12 (2000), 163-76.
Dixon, B., *Birth and death certificates. England and Wales 1837 to 1969*, (Burnham, 1999).
Dixon, B., *Marriages and certificates in England & Wales*, (Burnham, 2000).
Drake, M., 'Ashford 1840-1870: a socio-demography study', (unpublished Final Report, Centre for Research in the Social Sciences in the University of Kent at Canterbury, 1970).
Edvinsson, S., 'The Demographic Data Base at Umea University: a resource for historical studies', chapter 14 in Kelly Hall *et al*, *Handbook*, (2000) 231-248.
Fellegi, I. and Sunter, A., 'A theory for record linkage', *Journal of the American Statistical Association,* 64 (1969) 1138-1210.
Ferrie, E. ed., *Life at 33: the fifth follow-up of the National Child Development Study*, (London, 1993).
Ferrie, J. How ya gonna keep 'em down on the farm [when they've seen Schenechtady]'? Rural to urban migration in nineteenth-century America 1850-1870. Working paper, (Northwestern University, 1999).

Ferrie, J., 'Longitudinal data for the analysis of mobility in the U.S., 1850-1910' Available at: <http://faculty.econ.northwestern.edu/faculty/ferrie/papers/saltlakecity.pdf>.
Fogelvik, S. 'The Stockholm Historical Database at work', in Denley, P., Fogelvik, S. and Harvey, C. eds, *History and Computing II*, (Manchester, 1989) 256-65.
Fure, E., 'Interactive record linkage: the cumulative construction of life courses', *Demographic Research,* 3 (2000). Available at: www.demographic-research.org.
Geschwind, A. and Fogelvik, S., 'The Stockholm Historical Database', in Kelly Hall, McCaa, R. and Thorvaldsen, G. eds, *Handbook of International Historical Microdata for Population Research,* (Minneapolis, 2000) 207-229.
Guest, A., 'Notes from the National Panel Study: linkage and migration in the late nineteenth century', *Historical Methods,* 20 (1987) 63-77.
Gunn, P.A., 'The reprocessing of Tasmania's colonial censuses and vital registers: progress, problems and prospects'. [undated typescript, in library of Cambridge Group.]
Gutmann, M., Fliess, K., Holmes, A., Fairchild, A. and Teas, W., 'Keeping track of our treasures: managing historical data with relational database software'. *Historical Methods* 22 (1989) 128-43.
Hallas, C. S., 'The social and economic impact of a rural railway: the Wensleydale line', *Agricultural History Review*, 34 (1986) 29-44.
Hautaniemi, S. I., Anderton, D. L. and Swedlund, A., 'Methods and validity of a panel study using record linkage: matching death records to a geographic census sample in two Massachusetts towns, 1850-1912', *Historical Methods*, 33 (2000) 16-29.
Hendrickx, F., 'Nominal record linkage in a multi-lingual environment: two amendments to the Guth algorithm', *VGI Cahier,* 11 (1999) 121-35.
Hershberg, T., Burstien, A. and Dockhorn, R., 'Record Linkage', *Historical Methods Newsletter*, 2/3 (1976) 137-63.
Higgs, E., *Making sense of the census: the manuscript returns of the census 1801-1901*, (London, 1989).
Hood, D., 'Matching multiple data sources from New Zealand: the experience of the Caversham project', *History and Computing*, 12 (2000), 227-43.
Janssens, A., 'Managing longitudinal historical data: an example from nineteenth century Dutch population registers', *History and Computing,* 3 (1991) 161-74.
Katz, M. and Tiller, J., 'Record Linkage for Everyman: A Semi-Automatic Process', *Historical Methods Newsletter,* 5 (1972) 144-50.
Kelly, D., 'Linking nineteenth-century manuscript census records: a computer strategy', *Historical Methods Newsletter,* 7 (1974) 72-82.
Kitts, A., Doulton, D. and Reis, E. *The reconstruction of Viana do Castelo* (Egham, 1990).
Lait, A.J., and B. Randell, 'An assessment of name matching algorithms', http://homepages.cs.nc1.ac.uk/brian.Randell/home.informal/Genealogy/NameMatching.pdf , 14-08-03.
Lawton, R. 'Census data for urban areas', in Lawton, R. ed. *The census and social structure*, 82-145.
Lawton, R. ed., *The census and social structure: an interpretative guide to 19th century censuses for England and Wales*, (London, 1978).

Légaré, J., Lavoie, Y. and Charbonneau, H., 'The Early Canadian population: problems in automatic record linkage', *Canadian Historical Review,* 53 (1972) 427-42.
Leonard, S., Gutmann, M. and Sylvester, K., 'Demography and environment in grassland settlement: using linked longitudinal and cross-sectional data to explore household/agriculture systems'. [unpublished typescript provided to PI.]
Mandemakers, K., 'Historical Sample of the Netherlands (HSN). Background, objectives and international context.', in Marker, H. J. and Pagh, K. eds, *Yesterday. Proceedings from the 6th International Conference Association of History and Computing, Odense 1991*, (Odense, 1994) 174-81.
Mandemakers, K., 'The Netherlands. Historical Sample of the Netherlands', in Kelly Hall, McCaa, R. and Thorvaldsen, G. eds, *Handbook of International Historical Microdata for Population Research,* (Minneapolis, 2000) 149-77.
Mandemakers. K., 'The Historical Sample of the Netherlands', *Historical Social Research,* 26 (2001) 4, 179-90.
Miller, R. and Thovaldsen, G., 'Beyond record linkage: longitudinal analysis of turn-of-the-century inter-urban Swedish migrants', *History and Computing,* 9 (1997) 106-21.
Mineau, G. P., Bean, L. L. and Anderton, D. L., 'Description and evaluation of linkage of the 1880 census to family genealogies: implications for Utah fertility research', *Historical Methods,* 22 (1989) 144-57.
Mitch, D., 'Literacy and occupational mobility in rural versus urban Victorian England: evidence from the linked marriage register and census records for Birmingham and Norfolk, 1851 and 1881', *Historical Methods,* 38 (2005) 26-38.
Morris, R. J., 'Editorial: nominal record linkage into the 1990s', *History and Computing,* 4 (1992) iii-vii.
Morris, R. J., 'Qualitative to qualitative by way of coding and nominal record linkage. The search for the British middle class', *History and Computing*, 11 (1999) 9-29.
Morton, G., 'Presenting the Self: Record Linkage and Referring to Ordinary Historical Persons', *History and Computing,* 6 (1994) 12-20.
Newcombe, H. B., *Handbook of record linkage* (Oxford, 1988).
Nissel, M., *People count: a history of the General Register Office*, (London, 1987)
Pouyez, C., Roy, R. and Martin, F., 'The linkage of census name data: problems and procedures', *Journal of Interdisciplinary History*, 14 (1983) 129-52.
Reay, B., 'Before the transition: fertility in English villages, 1800-1880', *Continuity and Change*, 9 (1994) 91-120.
Richardson, S., 'Letter-cluster sampling and nominal record linkage', *History and Computing,* 6 (1994) 168-177.
Roberts, D. and Roberts, M., 'Surnames and relationships: an Orkney study', *Human Biology* 55 (1983) 341-7.
Rogers, H. J. and Willett, P., 'Searching for historical word forms in text databases using spelling-correction methods: reverse error and phonetic coding methods', *The Journal of Documentation,* 47 (1991) 333-53.
Royle, S. A., 'Irish manuscript census records: a neglected source of information', *Irish Geography*, 2 (1978) 110-25.
Russell, N. and Phelps, A., *Youth Cohort Study. Cohort 10 Sweep 1 (C10S1)*. Technical report (2001).

Ruusalepp, R., 'Nominal record linkage revisited: towards more interactive linking of records', paper presented at the XI AHC conference, Moscow, Aug 1996.
Scott, J. and Alwin, D., 'Retrospective versus prospective measurement of life histories in longitudinal research', in Stouthamer-Loeber, M. and van Kammen, W., eds, *Data Collection and Measurement: a practical guide,* (London, 1995), 98-127.
Shepherd, P., Smith, K., Joshi, H. and Dex, S., *Millennium Cohort Study first survey: a guide to the SPSS dataset*, Centre for Longitudinal Studies (London, 2003).
Smith, K. and Joshi, H. 'The Millennium Cohort Study', *Population Trends*, 107 (2002) 30-4.
Spencer, A. 'Scotlands people web site', *Computers in Genealogy*, 8 (2003), 55-7.
Stenflo, G. and Sundin, J., 'Using a large historical database. An example from the Demographic Database in Umeå', in Denley, P. and Hopkin, D. eds, *History and Computing*, (Manchester, 1987) 58-62.
Taylor, I. C. 'Liverpool's institutional and quasi-institutional populations in 1841 and 1851', in Mills and Schürer eds, *Local communities*, 42-6.
Tepping, B., 'A model for optimum linkage of records', *Journal of the American Statistical Association* 63 (1968) 1321-32.
Tilley, P. and French, C., 'Record Linkage for Nineteenth-Century Census Returns: Automatic or Computer-Aided?' *History and Computing,* 9 (1997) 123-33.
Vetter, J.E., Gonzalez, J. R. and Gutmann, M. P., 'Computer-assisted record linkage using a relational database system', *History and computing,* 4 (1992) 34-51.
Winchester, I., 'The linkage of historical records by man and computer', *Journal of Interdisciplinary History,* (1970) 107-24.
Winchester, I., 'What every historian needs to know about record linkage for the microcomputer era', *Historical Methods,* 25 (1992) 149-65.
Wrigley, E.A., (ed.), *Identifying people in the past* (London, 1973).
Zobel, J. and Dart, P., 'Phonetic string matching: lessons from information retrieval', (1996). Available at: <http://www.seg.rmit.edu.au/research/download.php?manuscript=105>.